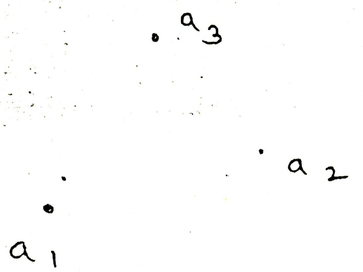
Finite Element Methods. General Remarks

The finite element method essentially is a numerical method for approximating the solution of second order or fourth order (or other types of) problems posed in variational form over a space V. A well-known approach for approximating such problems is Galerkin's method, that consists in defining similar problems, called discrete problems, over finite dimensional subspaces V_h of V. The finite element method in its simplest form is a Galerkin's method characterized by three basic aspects in the construction of V_h: (1) A triangulation $\mathcal{T}_h$ is established over the set $\bar{\Omega}$, i.e. $\bar{\Omega}$ is written as a union of finite elements $K \in \mathcal{T}_h$. (2) The functions $v_h \in V_h$ are piecewise polynomials; for each $K \in \mathcal{T}_h$ the spaces $P_K = \{ v_{h|K} ; v_h \in V_h \}$ consist of polynomials. (3) There should exist a basis in the space V_h whose functions have small supports. The degrees of freedom of a finite element are those parameters which uniquely define a function in the space P_K and consist of the values at a set of nodes. Let Σ_h be the

set of degrees of freedom. Then $\Sigma_K = \{ p(a_i),\ i = 1, \ldots n$ for finite elements K that are simplexes or polyhedra in R^n.

<u>Courant's triangle</u> or triangle of type 1. $(n=2)$

$\dim P_K = 3$.



The degrees of freedom may be point values (Lagrange finite elements) or some degrees of freedom are directional derivatives (Hermite finite elements).

X_h = a finite element space. It is associated with a triangulation $\mathcal{T}_h$ of a set $\bar{\Omega} = \cup_{K \in \mathcal{T}_h} K$

(FEM 1) and for each finite element $K \in \mathcal{T}_h$, the space P_K is

$$P_K = \{ v_{h|K} ;\ v_h \in X_h \}$$

For each $K \in \mathcal{T}_h$ the space P_K consists of polynomials or "nearly polynomials" (FEM 2)

There must exist one canonical basis in the space X_h whose functions have 'small' supports and are easy to describe (FEM 3)

Numerical Approximation

Let $[a,b]$ contain the $n+1$ points $x_0 < x_1 < \dots < x_n$ and $f, f', f'', \dots f^n$ exist and be continuous on $[a,b]$ and let $f^{(n+1)}$ exist on $a < x < b$. Then for any x on $[a,b]$, there exists a number ξ in (a,b) such that

$$f(x) - p_n(x) = (x - x_0) \dots (x - x_n)\ f^{(n+1)}(\xi) / (n+1)!$$

$p_n(x)$ being the polynomial of degree n (or less) that interpolates f at the points x_i $(i = 0, \dots, n)$.

(Lagrange interpolation).

$$p_n(x) = \sum_{i=0}^{n} f(x_i)\ L_i(x) \qquad (2)$$

and

$$L_i(x) = \prod_{\substack{j=0 \\ (j \neq i)}}^{n} \left(\frac{x - x_j}{x_i - x_j} \right) \quad (i = 0, 1, \dots, n) \qquad (3)$$

The best points $(x_0, \dots x_n$ to interpolate would be (since $f^{(n+1)}(\xi)$ involves the unknown ξ) such as to have

$$\max_{a \leq x \leq b} (x - x_0)(x - x_1) \dots (x - x_n) \qquad (4)$$

take its minimum possible value. By taking

$$x = \tfrac{1}{2}(b + a) + \tfrac{1}{2}(b - a)\, t$$

we have $a \leq x \leq b$ represented as $-1 \leq t \leq 1$

$$\inf_{p \in P_n} \max_{a \le x \le b} |f(x) - p(x)| = \inf_{p \in P_n} \max_{-1 \le t \le 1} |g(t) - p(t)|$$

where $g(t) = f\left(\frac{1}{2}(b-a)t + \frac{1}{2}(b+a)\right)$,

$$p(t) = p_n\left(\frac{1}{2}(b-a)t + \frac{1}{2}(b+a)\right)$$

The expression $\max_{-1 \le x \le 1} |(x-x_0)(x-x_1) \cdots (x-x_n)|$ attains its minimum value when the numbers x_i are the zeros of the Chebyshev polynomial (6)

$$T_{n+1}(x) = \cos\left[(n+1)\cos^{-1} x\right]$$

and

$$\inf_{p \in P_n} \max_{a \le x \le b} |f(x) - p(x)| = \frac{2}{(n+1)!}\left(\frac{b-a}{4}\right)^{n+1} |f^{n+1}(\xi)|$$

if $f^{(n+1)}(x)$ is continuous on $[a, b]$. In (7), $\xi \in (a, b)$.

Selection of Finite Element Methods

Engineering problems aiming to achieve quantitative information are tackled by seeking approximate solution(s of ~~a~~ mathematically formulated variants of the engineering problems in question and the approximations are achieved by the use of discretization methods that allow the determination within a range of accuracy.

H = linear space of possible solutions and u $(\in H)$ is the (mathematically) exact solution. $S \subset H$ is a finite dimensional linear subspace of H and may be called the trial space in which the approximate solution is sought. $B(u, v)$ is a bilinear form of u and v, $u \in H$ and $v \subset V$, V being a test space (of the same dimension as S, so that $B(u, v)$ is defined on $H \times V$. Pu = the approximate solution determined from the conditions

$$Pu \in S, \quad B(Pu, v) = B(u, v) \quad \forall v \in V \quad (1)$$

To compute Pu we must have the fulfilment of the following two conditions.

(1) For any $v \in V$, $B(u, v)$ is computable from the data of the problem, of course without the knowledge of u.

(2) For any $s \in S$, there is some $v \in V$ such that $B(s, v) \neq 0$ (2)

For any $u \in S$, $Pu = u$. The condition (1) leads to a set of linear equations that is uniquely solvable. The selection of S, V and B will determine the approximate solution Pu. We accept the approximation Pu if

$$\| Pu - u \| \leq \tau \| u \| \qquad (3)$$

τ being an a priori choice of the tolerance factor. (— we can have other types of criteria for the acceptance of an approximation; in particular, we may have an absolute error criterion as well). Thus, given the norm $\| \cdot \|$ and the tolerance factor τ, the aim is the selection of S, V and B so that (3) is realized in the most effective way.

We can have a variety of possible variational methods of discretization for a given problem.

A longitudinally loaded bar on an elastic support.

$u(x)$ = longitudinal displacement, $\sigma(x)$ = normal stress

$$E(x)\, u'(x) = \sigma(x), \qquad -\big(F(x)\,\sigma(x)\big)' + c(x)\, u(x) = p(x);$$

$$u(0) = 0, \qquad u(L) = 0 \qquad (4)$$

$E(x)$ = elasticity modulus, $F(x)$ = cross section area

$c(x)$ = spring constant, $p(x)$ = longitudinal load.

Variational formulations of the above problem.

$$B_1(u,\sigma;v,p) = \int_0^l (Eu'p - \sigma p + F\sigma v' + cuv)\,dx$$

$$= \int_0^l pv\,dx \qquad (5)$$

where $v \in \overset{\circ}{H}{}^1$, $p \in H^0$,

$$\overset{\circ}{H}{}^1 = \left\{ v;\ \int_0^l (v^2 + v'^2)\,dx < \infty,\ v(0) = v(l) = 0 \right\} \qquad (6)$$

$$H^0 = \left\{ p;\ \int_0^l p^2\,dx < \infty \right\} \qquad (7)$$

$B_1(u,\sigma;v,p)$ is thus computable without the prior knowledge of (u,σ) and it is defined for all $(u,\sigma) \in \overset{\circ}{H}{}^1 \times H^0 = H$; all $(v,p) \in H$.

$$\left.\begin{aligned} B_2(u,\sigma;v,p) &= \int_0^l pv\,dx \\ &= \int_0^l \left(u'p - \frac{\sigma p}{E} + F\sigma v' + cuv\right)dx \end{aligned}\right\} \qquad (8)$$

$$B_3(u,\sigma;v,p) = \int_0^l \left[-up' - \frac{\sigma p}{E} - (F\sigma)'v + cuv\right]dx$$

$$= \int_0^l pv\,dx \qquad (9)$$

Here $u, v \in H^0$, $\sigma, p \in H^1 = \left\{ v;\ \int_0^l (v^2 + v'^2)\,dx < \infty \right\}$

$$B_4(u,v) = \int_0^l (EFu'v' + cuv)\,dx \qquad (10)$$

$$= \int_0^l pv\,dx$$

Here $u, v \in \overset{\circ}{H}{}^1$.

The last [E(10)] is the usual form used in displacement finite element methods. — These can be infinitely many variational formulations —.

For a specific discretization method, apart from the bilinear form, we must also select the trial space S and the test space V. In the case of B, for instance, we may select finite dimensional subspaces S and V of H, both S and V having the same dimension and with $B(s, v) \neq 0$.

Norms. Let $u^{(k)} = \dfrac{d^k u}{dx^k}$ $(k = 0, 1, 2, \ldots)$

$$\|u\|_{H^k} = \left(\int_0^l \{ u^2 + u'^2 + \ldots + (u^{(k)})^2 \} \, dx \right)^{1/2} \quad (11)$$

Alternatively,
$$\|u\|_{H^k} = \sum_{j=0}^{k} \operatorname*{ess\,sup}_{0 \le x \le l} |u^{(j)}(x)| \quad (12)$$

$$\text{Let} \quad Z(u, S) = \inf_{s \in S} \|u - s\| \quad (13)$$

$Z(u, S)$ measures the error in the best possible approximation of u by elements of S with respect to the chosen norm $\|\cdot\|$. For acceptance of the approximation we must have $Z(u, S) \le \tau \|u\|$. (3)

The selection of S is essential to the effectiveness of the discretization method. u is not known a priori and selection of S must be such that every element

in H (or any other class to which u may belong) can be approximated well. More information about u allows more effective choice of S. This can be had through a learning process during the computation and S can be selected adaptively. For the approximation to be close to the exact solution, $Z(u,S)$ must be small and the method must be quasi-optimal. i.e

$$\|u - Pu\| \leq K\, Z(u,S) = K \inf_{s \in S} \|u - s\| \quad \forall u \in H \quad (14)$$

where K is a constant that is not too large. The smallest value of K for which (14) holds is called the <u>quasi-optimality constant</u>. If $\|Pu\| \leq K^* \|u\|$ for all $u \in H$, the smallest K^* for which the above holds is called the stability constant.

$$\|u - Pu\| = \|u - s - P(u-s)\| \leq \|u - s\| + \|P(u-s)\|$$

$$\leq \|u - s\| + K^* \|u - s\| = (K^* + 1)\|u - s\|$$

on using $Ps = s$ and hence K in (14) can be identified with $K^* + 1$; more specifically $K \leq K^* + 1$

Likewise $\|Pu\| \leq \|Pu - u\| + \|u\| \leq K\, Z(u,S) + \|u\|$

$$\leq (K+1)\|u\|$$

so that $K^* \leq K + 1$. The notions of quasi-optimality and stability are equivalent.

Approximability is affected only by the choice of the trial space S but stability and/or quasi-optimality depend on the interplay of B, H, S and V. Test functions belonging to the space V are not needed for approximation and the main aim of selection of V is to achieve ~~optimality~~ stability with the smallest possible K. Both approximability and stability depend on the norm used. For the rational selection of discretization procedures, there are other features besides approximability and stability, though these are of primary interest for the method.

Both approximability and stability depend on parameters that are present in the bilinear form B, e.g. E, F, η, p in the above example of a longitudinally loaded bar. When the performance of a method is relatively uninfluenced by the variation of the parameters within a wide range, the method is robust.

The acceptance criterion $\|u - Pu\| \leq \tau \|u\|$ in (3) above or $\|u - Pu\| \leq K\, Z(u, S)$ in (14) have no practical importance directly, since we do not know the quasi-optimality constant K or $Z(u, S)$. The quasi-optimality constant K is based on the worst

case since ~~u ∈ H in~~ (14) must hold for all $u \in H$.

A reliable acceptance criterion is to be based on a posteriori analysis of the approximate solution Pu. Let ϵ be a computable error estimator such that

$$c_1 \epsilon \leq \| u - Pu \| \leq c_2 \epsilon \tag{15}$$

and $$\theta = \frac{\epsilon}{\| u - Pu \|} \to 1 \quad \text{as} \quad \epsilon \to 0.$$

The quality of approximation obtained by finite element methods is assured by stability in conjunction with approximability and the stability of a method depends on the interplay between S, V and the bilinear form B besides the norm $\| \cdot \|$

<u>Example</u>: Longitudinally loaded bar. Take $l = 1$, $EF = 1$, $c = 0$ in B_4 (10).

$$B(u, v) = \int_0^1 u' v' \, dx \tag{16}$$

Solution $u \in \mathring{H}^1$ and $B(u, v) = \int_0^1 p v \, dx$ for all $v \in \mathring{H}^1$. The related two-point boundary value problem is

$$-u'' = p, \quad u(0) = u(1) = 0 \tag{17}$$

Let $\Delta = \{ 0 = x_0 < x_1 < \ldots < x_n = 1 \}$ be a mesh of $[0, 1]$ with $h_i = x_i - x_{i-1}$. For $r \geq 1$,

the mesh is r quasi-uniform if

$$\max_{i,j} \frac{h_i}{h_j} \leq r \tag{18}$$

(A weaker restriction: $\max_i \frac{h_i}{h_{i\pm 1}} \leq r$)

For the mesh Δ define the splines $M^r(\Delta)$ = the space of all piecewise polynomials of degree r with $r-1$ continuous derivatives. $M'(\Delta)$ = the space of continuous piecewise linear functions and $\mathring{M}'(\Delta)$ = the space of piecewise linear functions with zero boundary ~~conditions~~ values

<u>Theorem</u>. Let $S = V = \mathring{M}^1(\Delta)$ for an arbitrary partition Δ. Then, for all $u \in \mathring{H}^1$

$$\| Pu - u \|_{H^1} \leq K \inf_{s \in S} \| u - s \|_{H^1} \tag{19}$$

$$\| Pu - u \|_{H^1_\infty} \leq K \inf_{s \in S} \| u - s \|_{H^1_\infty} \tag{20}$$

$$\| Pu - u \|_{H^0_\infty} \leq K \inf_{s \in S} \| u - s \|_{H^0_\infty} \tag{21}$$

with K = a constant that is independent of Δ.

For any $c > 0$ and any mesh Δ there exists $u \in H^1$ such that

$$\| Pu - u \|_{H^0} \geq c \inf_{s \in S} \| u - s \|_{H^0} \tag{22}$$

The results (19) – (22) hold for the forms $B(u,v)$ in (16)

Take the more general case

$$B(u,v) \equiv \int_0^1 E\,u'v'\,dx \qquad (23)$$

where $0 < e_0 \leq E(x) \leq e_1 < \infty$. If $E(x)$ is sufficiently smooth, the theorem above holds without change with the constant K in (19), (20), (21) depending on e_0, e_1 and the maximum of the derivatives of $E(x)$. (19) holds with K depending only on e_0 and e_1 but (21) needs the differentiability restrictions on E. This is expressed by saying that the performance of the method is more robust with respect to the parameter $E = E(x)$ in the norm $\|\cdot\|_{H^1}$ than it is in the norm $\|\cdot\|_{H^0_\infty}$.

Next, let us take the problem (4) with $F=1$, $c=0$.

$$E\,u' = \sigma(x), \quad \sigma'(x) = -p, \quad u(0) = 0 = u(1) \qquad (2$$

Let $0 < e_0 \leq E(x) \leq e_1 < \infty$. Take $S = V = \mathring{M}^1(\Delta) \times M^0(\Delta)$

<u>Theorem</u>. Let Δ be an arbitrary mesh and let P be the bilinear form $B_2 = \int_0^1 \left(u'\rho - \frac{\sigma\rho}{E} + \sigma v'\right) dx$ as mentioned in (8). If S and V are chosen above, viz., $S = V = \mathring{M}^1(\Delta) \times M^0(\Delta)$, there exists a constant K depending only on e_0 and e_1

Such that

$$\|u - Pu\|_{H^1} + \|\sigma - \bar{P}\sigma\|_{H^0} \leq K \inf_{(\chi,\psi)\in S} \{\|u-\chi\|_{H^1} + \|\sigma-\psi\| \tag{25}$$

$$\|u - Pu\|_{H_\infty} + \|\sigma - P\sigma\|_{H_\infty} \leq K \inf_{(\chi,\psi)\in S} \{\|u-\chi\|_{H^0_\infty} + \|\sigma-\psi\|_{H^0_\infty}\} \tag{26}$$

$$\|\sigma - \bar{P}\sigma\|_{H^0} \leq K \inf_{\psi \in M^0(\Delta)} \|\sigma - \psi\|_{H^0} \tag{27}$$

$$\|\sigma - P\sigma\|_{H^0_\infty} \leq K \inf_{\psi \in M^0(\Delta)} \|\sigma - \psi\|_{H^0_\infty} \tag{28}$$

statements analogous to (27), (28) for the error $\|u - Pu\|$ are not true.

<u>Theorem</u>. Let $V(E)$ be the variation of the function E and let $\tilde{\chi}$ be defined by the conditions $(E\tilde{\chi}')' = 0$ on (x_{i-1}, x_i), $\tilde{\chi}(x_i) = \chi(x_i)$ $(i = 0, 1, \dots, n)$, χ being a $\mathring{M}^1(\Delta)$. Then for $i = 1, 2, \dots, n$,

$$\|u - Pu\|_{H^0_\infty} \leq K(e_0, e_1, V(E)) \inf_{\chi \in \mathring{M}^1(\Delta)} \{\|u-\chi\|_{H^0_\infty} + \|u - \bar{\chi}\|_{H^0_\infty}\} \tag{29}$$

and

$$\max_i |u(x_i) - Pu(x_i)| \leq K(e_0, e_1, V(E)) \inf_{\chi \in \mathring{M}^1(\Delta)} \|u - \tilde{\chi}\|_{H^0_\infty} \tag{30}$$

Finite Element Methods and Applications

In a paper of his in - Bulletin of the American Mathematical Society Vol. 49 (1943), 1-23 entitled " Variational methods for the solution of problems of equilibrium and vibrations" -, Richard Courant hit upon the idea of finite elements, then, a novel approach to the solution of boundary value problems. With the emergence of the computer and its role in the numerical solution of physical problems, the use of finite elements grew into a new discipline and in the last two decades, it has had an extraordinary resonance in the study and numerical analysis of solutions of boundary value problems in diverse fields of Physics and Engineering.

In a boundary value problem one seeks a function that satisfies a certain differential equation in a region Ω and conforms to specified conditions on the boundary ($\partial\Omega$) of Ω. Problems of this type have the property that the solution minimizes a certain (energy) functional f defined on some set V of functions, or is a stationary point

of the functional. The task of solving the boundary value problem is thus equivalent to the search of a function in the function space V that makes the functional f stationary. We have herein a variational version of the original boundary value problem.

Euler - Lagrange equation of the calculus of variati

Let $a \leq x \leq b$, $-\infty < r, s < \infty$ and $F(x, r, s)$ be a real function defined on $[a, b]$ and let $F \in C^2$. V is the class of functions $v \in C^2[a, b]$ with $v(a) = \alpha$, $v(b) = \beta$ and

$$f(u) = \int_a^b F[x, u(x), u'(x)]\, dx, \quad u \in V. \quad (1)$$

The set V is <u>not</u> a linear space. The function $u \in V$ is a stationary point of f if and only if

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) = 0, \quad a < x < b \quad (2)$$

E - L variational equation for the stationary property of $f(u)$ in (1) above.

e.g.

$$\left.\begin{aligned} -u''(x) &= r(x), \quad 0 < x < 1 \\ u(0) = 0, &\quad u(1) = 1 \end{aligned}\right\} \quad (3)$$

$$V = \{v \in C^2[0, 1],\ v(0) = 0;\ v(1) = 1\}$$

$$\text{Stationary } f(u) = \int_0^1 \left[\tfrac{1}{2}(u')^2 - r(x)\,u\right] dx, \ u \in V, \quad (4)$$

If $F(x, r, s, t)$ is a real function defined on $a \le x \le b$, $-\infty < r, s, t < \infty$ and F has continuous partial derivatives of order ≤ 3, and $V = \{ v \in C^4[a,b],\ v(a) = \alpha_0,\ v'(a) = \alpha_1,\ v(b) = \beta_0,\ v'(b) = \beta_1 \}$, we have

$$f(u) = \int_a^b F(x, u, u', u'')\,dx, \quad u \in V. \qquad (5)$$

Stationarity of f needs that

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial F}{\partial u''}\right) = 0, \quad a < x < b. \qquad (6)$$

x x x x

Natural and Essential Boundary conditions.

Take $V = \{ v \in C^2[a,b];\ v(a) = \alpha,\ v(b) = \beta \}$

$$f(u) = \int_a^b F[x, u(x), u'(x)]\,dx, \quad u \in V. \qquad (7)$$

$$f(u + \tau\eta) = \int_a^b F(x, u + \tau\eta, u' + \tau\eta')\,dx$$

and $f^{(1)}(u;\eta) = \left[\frac{\partial}{\partial \tau} f(u + \tau\eta)\right]_{\tau=0}$

$$= \int_a^b \left[\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right)\right]\eta\,dx + \left[\left(\frac{\partial F}{\partial u'}\right)\eta\right]_a^b \quad \forall \eta \in \tilde{V} \qquad (8)$$

where $\tilde{V} = \{ \eta;\ \eta = u - v,\ u, v \in V \}$ is the space of test functions or perturbation functions.

For stationarity of $f(u)$, $f^{(1)}(u;\eta) = 0$ for all $\eta \in C^2[a,b]$ and so if we choose $\eta(a) = \eta(b) = 0$, we get the E-L equation (2).

Since the expression in (8) is to vanish for the stationarity of u, we must have

$$\left[\frac{\partial F}{\partial u'} \eta \right]_a^b = 0 \quad \forall \eta \in \tilde{V}. \qquad (9)$$

If η is any $C^2[a,b]$ we must have

$$\frac{\partial F}{\partial u'} = 0 \quad \text{for } x = a, \; x = b. \qquad (10)$$

Thus every stationary point of f must satisfy the E-L equation (2) and also the natural boundary conditions $\frac{\partial F}{\partial u'} = 0$ for $x = a, b$.

Stationarity of $f(u) = \int_a^b F(x, u, u')\,dx$

Essential B.C.	B.C. on η	Natural B.C.
$u(a) = \alpha$ $u(b) = \beta$	$\eta(a) = 0$ $\eta(b) = 0$	$\left[\frac{\partial F}{\partial u'}\right]_{x=a} = 0$ $\left[\frac{\partial F}{\partial u'}\right]_{x=b} = 0.$

We may seek a solution of the E – L equati[on]

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) = 0, \quad a < x < b \qquad (11)$$

under boundary conditions that are neither essential nor natural for the functional

$$f(u) = \int_a^b F(x, u, u')\, dx .$$

Take $V = C^2[a, b]$, $f(u) = \int_a^b F(x, u, u')\, dx + [G(x,u)]_{x=a} + [H(x,u)]_{x=b}$, $u \in V$.

G and H have continuous first order partial derivatives.

Then $\tilde{V} = V$ and the E – L equation is the same (11) as before, if we take the natural boundary conditions in the form

$$\frac{\partial F}{\partial u'} - \frac{\partial G}{\partial u} = 0 \text{ for } x = a \text{ and } \frac{\partial F}{\partial u'} + \frac{\partial H}{\partial u} = 0 \text{ for } x = b. \qquad (12)$$

<u>Two – dimensional problems.</u>

Let $V = \{ v \in C^2(\bar{\Omega}),\ v = \alpha \text{ on } \Gamma = \partial\Omega \}$ (13)

$$f(u) = \iint_\Omega F(x, y, u, u_x, u_y)\, dx\, dy, \quad u \in V \qquad (14)$$

Condition for a stationary point of f is

$$\iint_\Omega \left\{ \frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) \right\} \eta\, dx\, dy$$

$$+ \int_\Gamma \left(\nu_1 \frac{\partial F}{\partial u_x} + \nu_2 \frac{\partial F}{\partial u_y} \right) \eta\, ds = 0$$

where (ν_1, ν_2) are the direction cosines of the outer normal to Γ. The E – L equation is

$$\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) = 0 \quad \text{in } \Omega \tag{15}$$

An essential boundary condition is $u = \alpha$ on Γ, and the natural boundary condition is

$$\nu_1 \frac{\partial F}{\partial u_x} + \nu_2 \frac{\partial F}{\partial u_y} = 0 \quad \text{on } \Gamma. \tag{16}$$

These two forms of boundary conditions are conforming respectively to the Dirichlet and Neumann types.

x x x x

Three-dimensional problems.

Condition for a stationary point of the functional

$$I = \iiint_{\Omega} F(x, y, z, u, u_x, u_y, u_z)\, dx\, dy\, dz, \quad u \in V \tag{17}$$

is that

$$\iiint_{\Omega} \left\{ \frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) - \frac{\partial}{\partial z}\left(\frac{\partial F}{\partial u_z}\right)\right\} \eta\, dx\, dy\, dz$$

$$+ \iint_{\Gamma} \left(\nu_1 \frac{\partial F}{\partial u_x} + \nu_2 \frac{\partial F}{\partial u_y} + \nu_3 \frac{\partial F}{\partial u_z}\right) \eta\, dS = 0 \tag{18}$$

for all $\eta \in \tilde{V}$.

Essential boundary condition: $u = \alpha(x, y, z)$ on Γ

Natural boundary condition: $\nu_1 \frac{\partial F}{\partial u_x} + \nu_2 \frac{\partial F}{\partial u_y} + \nu_3 \frac{\partial F}{\partial u_z} = 0$ on Γ.

x x x x

The Hilbert space $L_2(a,b)$; concept of completion.

$L_2[a,b]$ = the set of all integrable real functions $u(x)$, $-\infty < a \leq x \leq b < \infty$ such that $\int_a^b (u(x))^2 \, dx < \infty$. The set $L_2[a,b]$ constitutes a linear space since the conditions $u \in L_2[a,b]$, $v \in L_2[a,b]$ imply that $\alpha u + \beta v \in L_2[a,b]$ for all $\alpha, \beta \in R$.

$(u,v) = \int_a^b uv \, dx$, $u, v \in L_2[a,b]$

$\|u\| = \sqrt{(u,u)}$.

If $u, v, w \in L_2[a,b]$ and $\alpha, \beta \in R$,

$(u,v) = (v,u)$, $(\alpha u + \beta v, w) = \alpha(u,w) + \beta(v,w)$,

$(u,u) \geq 0$, $(u,u) = 0 \iff u = 0$,

$\|\alpha u\| = |\alpha| \|u\|$, $\|u+v\| \leq \|u\| + \|v\|$

$\|u\| \geq 0$, $\|u\| = 0 \iff u = 0$.

(u,v): inner product, $\|u\|$: norm

$|u,v| \leq \|u\| \|v\|$ for all $u, v \in L_2[a,b]$.

A sequence $\{u_n\} \subset L_2[a,b]$ converges in the mean to $u \in L_2[a,b]$ if $\lim_{n\to\infty} \|u - u_n\| = 0$.

Cauchy sequence: $\lim_{m,n\to\infty} \|u_m - u_n\| = 0$. (19)

If a sequence converges in the mean, it is a Cauchy sequence. For $L_2[a,b]$, the converse statement is also true and the space $L_2[a,b]$ is complete on account of this property. A linear space equipped with an inner product and norm $\|u\| = \sqrt{(u,u)}$ and with validity of the completeness property is a Hilbert space.

A linear functional on $L_2[a,b]$ is a mapping $\phi : L_2[a,b] \to R$ such that $u, v \in L_2[a,b]$ with $\alpha, \beta \in R$ implies that $\phi(\alpha u + \beta v) = \alpha \phi(u) + \beta \phi(v)$. A functional is bounded if there exists $\alpha > 0$ such that

$$|\phi(u)| \leq \alpha \|u\| \quad \forall\, u \in L_2[a,b]. \quad (20$$

<u>Riesz Representation Theorem</u>. Let ϕ be a bounded linear functional on $L_2[a,b]$. Then there exists a unique $u_0 \in L_2[a,b]$ such that

$$\phi(u) = (u, u_0) \quad \forall\, u \in L_2[a,b].$$

<u>A simple example</u>

Let $V = \{ v \in C^2[a,b] ;\ v(a) = v(b) = 0 \}$ (2

$u \in V$, $f(u) = \int_a^b \left[\frac{1}{2} p(x) (u')^2 + \frac{1}{2} q(x) u^2 - g(x) u \right] dx$ (2

where $p \in C^1[a,b]$, $q, g \in C[a,b]$; $0 < p_0 \leq p(x) \leq p$

$0 \leq q(x) \leq q_1$ for $a \leq x \leq b$, (p_0, p_1, q_1) are constants. The corresponding set of test functions is

$$\tilde{V} = \{ \eta ; \eta = u - v, \quad u, v \in V \}$$

and clearly $\tilde{V} \equiv V$. Introduce

$$a(u, v) = \int_a^b \{ p(x) u' v' + q(x) u v \} dx, \quad u, v \in V \qquad (23)$$

$$G(u) = \int_a^b g(x) u \, dx, \quad u \in V. \qquad (24)$$

Then we have

$$f = \tfrac{1}{2} a(u, u) - G(u), \quad u \in V$$

G is a linear mapping of V onto R, a is a bilinear mapping of $V \times V$ onto R; further a is symmetric bilinear form and

$$a(u + \eta, u + \eta) = a(u, u) + 2 a(u, \eta) + a(\eta, \eta)$$

$$G(u + \eta) = G(u) + G(\eta).$$ Hence we identify

$$f(u + \eta) = f(u) + [a(u, \eta) - G(\eta)] + \tfrac{1}{2} a(\eta, \eta)$$

for all $u, \eta \in V$. Let $\tau \in R$, $\|\eta\| = 1$,

$$f(u + \tau \eta) = f(u) + \tau [a(u, \eta) - G(\eta)] + \tfrac{1}{2} \tau^2 a(\eta, \eta)$$

If $f^{(m)}(u; \eta) = \left[\frac{d^m}{d\tau^m} f(u + \tau \eta) \right]_{\tau = 0}$

$$f^{(1)}(u; \eta) = a(u, \eta) - G(\eta), \qquad f^{(2)}(u; \eta) = a(\eta, \eta)$$

$f^{m}(u; \eta) = 0 \quad (m = 3, 4, \cdots)$

u is a stationary point of f if $f^{(1)}(u; \eta) = 0$ $\forall \eta \in V, \|\eta\| = 1$. i.e.

$$a(u, \eta) = G(\eta), \quad \forall \eta \in V \tag{25}$$

Let $\mathcal{L}$ be the mapping of V onto $C[a,b]$ defined by

$$\mathcal{L}u = -(p(x)u')' + q(x)u, \quad u \in V \tag{26}$$

The operator $\mathcal{L}$ (i.e. a mapping of one set of functions into another set) is linear and

$$a(u, \eta) = \int_a^b (\mathcal{L}u)\eta \, dx, \quad u, \eta \in V. \tag{27}$$

i.e. $a(u, \eta)$ = the inner product $(\mathcal{L}u, \eta)$

$G(\eta) = (g, \eta)$

$a(u, \eta) = G(\eta) \quad \forall \eta \in V$ implies $(\mathcal{L}u - g, \eta) = 0$ for all $\eta \in V$ and hence $\mathcal{L}u = g$.

The E-L equation for f in (22) is $\mathcal{L}u = g$ and the essential boundary conditions are $u(a) = u(b) = 0$.

$$f(u) = f(\hat{u}) + \frac{1}{2} a(u - \hat{u}, u - \hat{u}) \tag{28}$$

and $\min_{u \in V} f(u) = f(\hat{u})$ where $\hat{u}$ is the unique solution of the differential equation

$\mathcal{L}u = g, \quad u(a) = u(b) = 0.$

Enlargement of the set of admissible functions.

Domain of definition of f extended from the set V into a larger set $\bar{V}$. such that

(1) every function in $\bar{V}$ can be approximated arbitrarily closely (in some accepted sense) by functions in V

(2) the solution $\hat{u}$ of the two-point boundary value problem $\mathcal{L}u = g$, $u(a) = u(b) = 0$ is the unique solution of the variational problem of finding $u \in \bar{V}$ such that $a(u, \eta) = G(\eta)$ for all $\eta \in \bar{V}$

(3) $\min_{u \in V} f(u) = \min_{u \in \bar{V}} f(u) = f(\hat{u})$.

$L_2[a,b]$ is complete with respect to the norm $\|u\| = \sqrt{(u,u)}$. A Cauchy sequence in the norm converges to $u \in L_2[a,b]$ in the mean in the norm

Let $(u, v)_{\mathcal{L}} = a(u, v) = \int_a^b [p(x)\, u' v' + q(x)\, u v]\, dx$

for all $u, v \in V$, $\|u\|_{\mathcal{L}} = (u,u)_{\mathcal{L}}^{1/2} = (a(u,u))^{1/2}$

~~for~~ $u \in V$. The above inner product and norm are well-defined on V.

$(\cdot,\cdot)_{\mathcal{L}}$ = energy inner product

$\|\cdot\|_{\mathcal{L}}$ = energy norm

The space V is not complete in the energy norm.

Let $\{u_n\} \subset V$ be a Cauchy sequence in the norm $\|\cdot\|_{\mathcal{L}}$. $\lim_{m,n\to\infty} \|u_n - u_m\|_{\mathcal{L}} = 0$.

$$\|u\|^2_{\mathcal{L}} = a(u,u) = (pu', u') + (qu, u) \geq p_0 (u', u')$$

(since $0 < p_0 \leq p(x) \leq p_1$, $0 \leq q(x) \leq q_1$)

$$\|u\|^2_{\mathcal{L}} \geq \frac{p_0}{2}(u', u') + \frac{p_0}{2}(u', u') \geq \frac{p_0}{2}(u', u') + \frac{p_0}{(b-a)^2}(u$$

(since $(u', u') \geq \frac{2(u,u)}{(b-a)^2}$ for all $u \in \{v \in C^1[a,b]$

$v(a) = 0\}$. Let $\rho = p_0 \min\{\frac{1}{2}, (b-a)^{-2}\} > 0$

$$\|u\|^2_{\mathcal{L}} \geq \rho(\|u'\|^2 + \|u\|^2) \quad \forall u \in V. \qquad (31$$

For a Cauchy sequence in the norm $\|\cdot\|_{\mathcal{L}}$,

$\lim_{m,n\to\infty} \|u'_m - u'_n\| = 0$ in the norm $\|\cdot\|$.

Since $L_2[a,b]$ is a complete space, there exist elements $u, \tilde{u}$ in $L_2[a,b]$ such that

$$\lim_{n\to\infty} \|u - u_n\| = 0, \quad \lim_{n\to\infty} \|\tilde{u} - u'_n\| = 0. \qquad (32$$

Define $\bar{V}$ to be the set of all functions u that can be obtained by this process.

$\tilde{u}$ = the generalized derivative of u and is uniquely determined by u. If $u \in C^1[a,b]$, $\tilde{u} = u'$.

The space $\bar{V}$ is complete with respect to the energy norm and is therefore a Hilbert space.

Sobolev Spaces (one dimension)

Let $(u, v)_1 = \sum_{i=0}^{1} (u^{(i)}, v^{(i)}) = \int_a^b [uv + u'v']\,dx$, $u, v \in V$ (

$$\|u\|_1 = (u, u)_1^{1/2} = \left\{ \int_a^b (u^2 + u'^2)\, dx \right\}^{1/2}, \quad u \in V \quad (3$$

where $V = \{ v \in C^2[a, b] ;\ v(a) = v(b) = 0 \}$

$(\cdot,\cdot)_1$, $\|\cdot\|_1$ are well-defined inner product and norm respectively and are the Sobolev inner product and Sobolev norm of order one. The energy norm

$$\|u\|_L = \left\{ \int_a^b (p u'^2 + q u^2)\, dx \right\}^{1/2}, \quad u \in V \qquad (35)$$

and the Sobolev norm $\|\cdot\|_1$ are equivalent in the sense that constants $\rho, \beta > 0$ exist such that

$$\rho \|u\|_1^2 \le \|u\|_L^2 \le \beta \|u\|_1^2 \quad \forall\, u \in V$$

$\beta = \max(p_1, q_1)$, $\rho = p_0 \min\left(\frac{1}{2}, (b-a)^{-2}\right)$.

The Sobolev inner product and norm are defined on all of $\bar{V}$ when first order derivatives are interpreted in the generalized sense and $\bar{V}$ is also a Hilbert space with respect to this inner product and norm.

Let $K = 0, 1, 2, \ldots$ and $C^K(a,b)$ = the set of functions defined on $a < x < b$ with continuous derivatives of order upto and including K.

$$\tilde{C}^K(a,b) = \left\{ v \in C^K(a,b) ; \int_a^b (v^{(i)})^2 dx < \infty , i = 0, 1, \ldots \right.$$

$\tilde{C}^K(a,b)$ is a proper subset of $C^K(a,b)$. Let

$$(u, v)_K = \sum_{i=0}^{K} (u^{(i)}, v^{(i)}) = \int_a^b (uv + u'v' + \ldots + u^{(K)} v^{(K)}) dx$$

$$\{u, v\} \in \tilde{C}^K(a,b).$$

$$\|u\|_K = (u,u)_K^{1/2} = \left\{ \int_a^b (u^2 + u'^2 + \ldots + (u^{(K)})^2) dx \right\}^{1/2}$$

$$u \in \tilde{C}^K(a,b).$$

The completion of $\tilde{C}^K(a,b)$ with respect to $\|\cdot\|_K$ produces a larger space $H^{(K)}(a,b)$ that may be called a Sobolev space. The domain of definition of the Sobolev inner product and norm is extended to all of $H^{(K)}(a,b)$.

<u>Function of compact support.</u> u defined on (a,b) has compact support on (a,b) if α, β exist such that $a < \alpha < \beta < b$ and $u(x) = 0$ for $a < x < \alpha$ and $\beta < x < b$. $\mathring{H}^K(a,b)$ = completion with respect to $\|\cdot\|_K$ of the subset of $C^K(a,b)$ with compact support on (a,b). This is also a Sobolev space and is a Hilbert space with respect to the Sobolev inner product and norm.

$$\mathring{H}^K(a,b) = \{ v \in H^K(a,b) ; v^{(i)}(a) = v^{(i)}(b) = 0 , i = 0, \ldots K-1$$

$$\bar{V} = \mathring{H}^1(a,b)$$

The subspace $C^1(a,b)$ with compact support on (a,b) is not the same as $V = \{ v \in C^2[a,b] ; v(a) = v(b) = 0 \}$ but their completions with respect to the norm $\| \cdot \|_1$ (i.e. $\|u\|_1 = (u,u)_1^{1/2} = \{ \int_a^b (u^2 + u'^2)\, dx \}^{1/2}$, $u \in \tilde{C}^1(a,b)$) are the same, i.e. $\bar{V} = \mathring{H}^1(a,b)$. The boundary value problem $\mathcal{L}u = g$, $u(a) = u(b) = 0$ associated with

$$f(u) = \int_a^b \left[\frac{1}{2} p(x) u'^2 + \frac{1}{2} q(x) u^2 - g(x) u\right] dx, \quad u \in V$$

has the formulation:

find $u \in \mathring{H}^1(a,b)$ such that $a(u,\eta) = G(\eta) \quad \forall\, \eta \in \mathring{H}^1(a,b)$

Finite Element solution of Boundary Value Problems

Hilbert Space $L_2[a,b]$.

Let $u(x)$ be an integrable real function over $[a,b]$ with $-\infty < a \leq x \leq b < \infty$ and $\int_a^b u^2(x)\,dx < \infty$

$L_2[a,b]$ = the set of all such functions $u(x)$.

If u, v are two arbitrary members of $L_2[a,b]$, the product uv is integrable and $\alpha u + \beta v \in L_2[a,b]$ for arbitrary $\alpha, \beta \in R$ so that $L_2[a,b]$ is a linear space of functions on (a,b)

$$(u,v) = \int_a^b uv\,dx, \quad \|u\| = \sqrt{(u,u)} \qquad (1)$$

We have for $\{u, v, w \in L_2[a,b]; \ \alpha, \beta \in R\}$

$$(\alpha u + \beta v, w) = \alpha(u,w) + \beta(v,w) \qquad (2)$$

$$(u,v) = (v,u) \qquad (3)$$

$$(u,u) \geq 0, \quad (u,u) = 0 \iff u = 0 \qquad (4)$$

$$\|\alpha u\| = |\alpha|\,\|u\|, \quad \|u+v\| \leq \|u\| + \|v\| \qquad (5)$$

$$\|u\| \geq 0, \quad \|u\| = 0 \text{ iff } u = 0 \qquad (6)$$

(u,v) = the inner product of u, v, $\|u\|$ is the norm of u. If $H(x)$ is a continuous

positive function on $a \leq x \leq b$, we may define the inner product and norm in the form

$$(u, v)_H = \int_a^b H(x)\, u\, v\, dx, \quad \|u\|_H = (u, u)_H^{1/2}. \quad (7)$$

If $u, v \in L_2[a, b]$, we have the Schwarz inequality

$$|(u, v)| \leq (\|u\|)(\|v\|) \quad (8)$$

<u>Mean convergence of a sequence</u> If $\{u_n\}$ is a sequence belonging to $L_2[a, b]$ such that $\lim_{n \to \infty} \|u - u_n\| = 0$ where u also belongs to $L_2[a$ then the sequence $\{u_n\}$ converges in the mean to u.

If $\lim_{m,n \to \infty} \|u_m - u_n\| = 0$, the sequence $\{u_n\}$ is a Cauchy sequence.

If a sequence $\{u_n\}$ converges in the mean to u, $\{u_n\}$ is necessarily a Cauchy sequence for

$$\|u_m - u_n\| = \|u_m - u + u - u_n\|$$
$$\leq \|u_m - u\| + \|u - u_n\| \quad (9)$$

For the space $L_2[a, b]$ the converse is also true; i.e. every Cauchy sequence in $L_2[a, b]$ is also a sequence that converges in the mean to some u.

The space $L_2[a,b]$ is "complete".

A linear space equipped with an inner product and norm such that $\|u\| = \sqrt{(u,u)}$ and with the property of completeness, is a Hilbert space.

$L_2[a,b]$ is a Hilbert space.

A linear functional on $L_2[a,b]$ is a mapping $\phi : L_2[a,b] \to R$ such that $\{u, v \in L_2[a,b],\ \alpha, \beta \in R\}$ implies

$$\phi(\alpha u + \beta v) = \alpha\,\phi(u) + \beta\,\phi(v) \qquad (10)$$

If $|\phi(u)| \leq \alpha \|u\| \ \forall\ u \in L_2[a,b]$ with $\alpha > 0$, the functional is bounded.

If u_0 is a fixed element of $L_2[a,b]$,

$$\phi(u) = (u, u_0), \qquad u \in L_2[a,b] \qquad (11)$$

is a bounded linear functional.

The theorem below (called 'Riesz Representation Theorem') shows that every bounded linear functional of $L_2[a,b]$ has a representation of the type (11), i.e. it is an innerproduct in $L_2[a,b]$ with one constituent equal to a fixed element of $L_2[a,b]$.

Theorem (Riesz) Let ϕ be a bounded linear

functional on $L_2[a,b]$. Then there exists a unique $u_0 \in L_2[a,b]$ such that

$$\phi(u) = (u, u_0) \qquad \forall \ u \in L_2[a,b].$$

This theorem is extendable to bounded linear functionals on any Hilbert space.

Concept of completion.

Let

$$f(u) = \int_a^b \left[\tfrac{1}{2} p(x)(u')^2 + \tfrac{1}{2} q(x) u^2 - g(x) u\right] dx, \quad (12)$$

with $u \in V$, $V = \{ v \in C^2[a,b];\ v(a) = v(b) = 0\}$,

$p \in C^1[a,b]$, $q, g \in C[a,b]$ and

$0 < p_0 \le p(x) \le p_1$, $0 \le q(x) \le q_1$

for $a \le x \le b$. p_0, p_1 are positive constants and q_1 is a non-negative constant. Let $\tilde{V}$ be the set $\tilde{V} = \{ \eta,\ \eta = u - v,\ u, v \in V\}$

$\tilde{V}$ is the space of test functions or perturbation functions corresponding to V and in the present case $\tilde{V} = V$. Introduce

$$a(u,v) = \int_a^b [p(x) u' v' + q(x) u v]\, dx, \qquad u, v \in V \quad (13)$$

$$G(u) = \int_a^b g(x) u \, dx, \qquad u \in V \quad (14)$$

The functional f in (12) is then expressible as

$$f = \tfrac{1}{2} a(u,u) - G(u), \qquad u \in V. \quad (15)$$

(7)

The mapping $G : V \to R$ is a linear functional on V and the mapping $a : V \times V \to R$ is bilinear and symmetric.

$$a(u, v) = a(v, u) \qquad u, v \in V \quad \text{(symmetry)}$$

$$\left.\begin{aligned} a(\alpha u + \beta v, w) &= \alpha\, a(u, w) + \beta\, a(v, w) \\ a(w, \alpha u + \beta v) &= \alpha\, a(w, u) + \beta\, a(w, v) \end{aligned}\right\} \qquad (17)$$

If $\tau \in R$ and $\eta \in V$ and $\|\eta\| = 1$

$$f(u + \tau\eta) = f(u) + \tau\left[a(u, \eta) - G(\eta)\right] + \frac{1}{2}\tau^2 a(\eta, \eta) \qquad (18)$$

Let $f^{(m)}(u ; \eta) = \left[\frac{d^m}{d\tau^m} f(u + \tau\eta)\right]_{\tau = 0}$

$f^{(1)}(u; \eta) = a(u, \eta) - G(\eta)$, $\quad f^{(2)}(u; \eta) = a(\eta, \eta)$,

$f^{(m)}(u; \eta) = 0 \qquad (m = 3, 4, \ldots)$

Thus u is a stationary point of f if

$$a(u, \eta) = G(\eta) \qquad \forall\, \eta \in V \qquad (19)$$

Let $\mathcal{L} u = -(p(x)\, u')' + q(x)\, u, \quad u \in V \qquad (20)$

$\mathcal{L}$ is a linear mapping of one set of functions of V onto another set; i.e. $\mathcal{L}$ is a linear operator and it is easily seen that

$$a(u, \eta) = \int_a^b (\mathcal{L} u)\, \eta \, dx, \quad u, \eta \in V. \qquad (21)$$

$$= (\mathcal{L} u, \eta)$$

$$G(\eta) = (g, \eta) \qquad (22)$$

The E-L equation for $f(u)$ in (12) is (19)

$\mathcal{L}u = g$ and $a(u,\eta) = G(\eta) \quad \forall\, \eta \in V$

is the variational formulation of the two-point boundary value problem $\mathcal{L}u = g$, $u(a) = u(b) = 0$.

This problem (in differential equations) has a unique solution $\hat{u}$ in V. Further, we have from (18),

$$f(u) = f(\hat{u}) + \tfrac{1}{2}\, a(u-\hat{u},\, u-\hat{u}) \qquad (23)$$

and since $a(\eta,\eta) > 0$ for all $\eta \in V$ with $\eta \neq 0$, we have

$$\min_{u \in V} f(u) = f(\hat{u}) \qquad (24)$$

* * * *

In the above the domain of definition of the functional f is the set V and it is enlarged into the set $\bar{V}$ such that

(1) every function in $\bar{V}$ can be approximated arbitrarily closely — in some well-defined sense — by functions in V

(2) the solution $\hat{u}$ of the two point boundary value problem $\mathcal{L}u = g$, $u(a) = 0 = u(b)$ is the unique solution of the variational problem of finding $u \in \bar{V}$ such that $a(u,\eta) = G(\eta) \quad \forall\, \eta \in \bar{V}$ (25)

(3) $\displaystyle\min_{u \in V} f(u) = \min_{u \in \bar{V}} f(u) = f(\hat{u})$ (26)

Let $(u, v)_{\mathcal{L}} = a(u, v) \quad \forall\, u, v \in V$ (27)

$$\|u\|_{\mathcal{L}} = (u,u)_{\mathcal{L}}^{1/2} = \{a(u,u)\}^{1/2}, \quad u \in V \qquad (28)$$

In (27), $a(u,v) = \int_a^b [p(x)u'v' + q(x)uv]\,dx$ (13)

$(\,,\,)_{\mathcal{L}}$ and $\|\cdot\|_{\mathcal{L}}$ are a well-defined inner product and norm on V and these may be called the energy inner product and energy norm respectivel[y]

The space V is not complete in the energy norm

Let $\{u_n\} \subset V$ be a sequence with the Cauchy property, viz.,

$$\lim_{m,n\to\infty} \|u_m - u_n\|_{\mathcal{L}} = 0.$$

$$\|u\|_{\mathcal{L}}^2 = (pu', u') + (qu, u) \geq p_0(u', u') = \frac{p_0}{2}\{(u',u') + (u',u')\}$$

$$\geq \frac{p_0}{2}(u', u') + \frac{p_0}{(b-a)^2}(u,u)$$ $\{$ We have $\forall\, u \in \{v \in C'[a,b],\ v(a) = 0\}$ $(u', u') \geq 2(b-a)^{-2}(u,u)$ [Poincare inequality].

$$\|u\|_{\mathcal{L}}^2 \geq \rho\left(\|u'\|^2 + \|u\|^2\right) \quad \forall\, u \in V$$

with $\rho = p_0 \min\{\frac{1}{2}, (b-a)^{-2}\} > 0$. (29)

$$\|u_m - u_n\|_{\mathcal{L}}^2 \geq \rho\left(\|u_m - u_n\|^2 + \|u'_m - u'_n\|^2\right) \qquad (3[0])$$

$$\lim_{m,n\to\infty} \|u_m - u_n\|^2 = 0, \quad \lim_{m,n\to\infty} \|u'_m - u'_n\| = 0$$

Since $L_2[a,b]$ is a complete space with respect to

the norm $\|\cdot\|$, elements $u, \tilde{u} \in L_2[a,b]$ exist such that

$$\lim_{n\to\infty} \| u - u_n \| = 0, \qquad \lim_{n\to\infty} \| \tilde{u} - u'_n \| = 0. \qquad (31)$$

$\bar{V}$ = the set of all functions u that can be defined by this process.

$\tilde{u}$ = generalized derivative of u and is uniquely determined by u. If $u \in C'[a,b]$, $\tilde{u} = u'$.

$$(u,v)_L = \int_a^b [p(x)u'v' + q(x)uv]\,dx, \qquad u, v \in \bar{V} \qquad (32)$$

$$\|u\|_L = (u,u)_L^{1/2}, \qquad u \in \bar{V} \qquad (33)$$

The space $\bar{V}$ is complete with respect to the energy norm $\|\cdot\|_L$ and is a Hilbert space.

$\bar{V}$ contains functions not contained in V, not contained even in $C'[a,b]$. It includes all continuous functions which are piecewise differentiable such that $u(a) = u(b) = 0$ and

$$\int_{x_{i-1}}^{x_i} [u'(x)]^2\,dx < \infty \quad \text{for } i = 1, 2, \ldots n+1$$

where $x_0 = a < x_1 < x_2 < \ldots \ldots < x_n < b = x_{n+1}$ and $u'(x)$ exists and is continuous everywhere in $[a,b]$ except at the points $x_1, x_2, \ldots, x_n$.

Essential boundary conditions imposed on V are carried over to $\bar{V}$ by the process of completion.

However, completing the space $W = \{v \in C^2[a,b];\ v(a) = 0,\ v'(b) = 0\}$ wherein the boundary condition $u'(b) = 0$ is natural and not essential, we have $u \in \overline{W}$ with $u'(b) \neq 0$. A natural boundary condition cannot be imposed during the process of completion.

x x x x x

SOBOLEV SPACES (ONE SPACE DIMENSION)

$$(u, v)_1 = \sum_{i=0}^{1} (u^{(i)}, v^{(i)}) = \int_a^b (uv + u'v')\,dx, \quad u, v \in V$$

$$\|u\|_1 = (u, u)_1^{1/2} = \left(\int_a^b (u^2 + u'^2)\,dx \right)^{1/2}, \quad u \in V$$

$$V = \{ v \in C^2[a,b];\ v(a) = v(b) = 0 \} \qquad (3\ldots$$

$(\,,\,)_1$ and $\|\cdot\|_1$ are well-defined inner product and norm respectively, on V and are called the Sobolev inner product and the Sobolev norm of order one.

The energy norm

$$\|u\|_{\mathcal{L}} = \left\{ \int_a^b (p\, u'^2 + q u^2)\,dx \right\}^{1/2}, \quad u \in V \qquad (37)$$

and the Sobolev norm $\|u\|_1$ are equivalent in the sense that constants $\rho, \beta > 0$ exist such that

$$\rho \|u\|_1^2 \leq \|u\|_{\mathcal{L}}^2 \leq \beta \|u\|_1^2 \quad \forall u \in V. \qquad (38)$$

Completion of V with respect to the energy norm $\|\cdot\|_{\mathcal{L}}$ leads to the extended space $\bar{V}$. We can similarly complete V with respect to the Sobolev norm $\|\cdot\|_1$ and arrive at the extended space $\tilde{V}$. For every sequence $\{u_n\} \subset V$, the Cauchy sequency property w.r.t. the Sobolev norm, viz., $\lim_{m,n\to\infty} \|u_n - u_m\|_1 =$ implies the Cauchy sequence property in the energy norm, viz., $\lim_{m,n\to\infty} \|u_n - u_m\|_{\mathcal{L}} = 0$ and conversely, in view of the result (38) above. We therefore arrive at the result that $\tilde{V} = \bar{V}$.

Sobolev spaces of higher orders.

K = a non-negative integer $(= 0, 1, 2, \dots)$

$C^K(a,b)$ = set of functions defined on $a < x < b$ with continuous derivatives of order upto and includin K.

$$\tilde{C}^K(a,b) = \left\{ v \in C^K(a,b) ; \int_a^b (v^{(i)})^2 \, dx < \infty \quad (i = 0, 1, \dots, K) \right\} . \quad (3$$

$C^K(a,b)$ includes functions that are unbounded near a or b and for which $\int_a^b v^2 \, dx = \infty$. e.g. $v(x) = (x-a)^{-1}$

$\tilde{C}^K(a,b)$ is therefore a proper subset of $C^K(a,b)$.

$$(u,v)_K = \sum_{i=0}^{K} (u^{(i)}, v^{(i)}) = \int_a^b [uv + u'v' + \dots + u^{(K)} v^{(K)}] \, dx, \quad (4$$

$u, v \in \tilde{C}^K(a,b)$

$$\|u\|_K = (u,u)_K^{1/2} = \left(\int_a^b \{u^2 + (u')^2 + \cdots + (u^{(K)})^2\} dx\right)^{1/2}, \quad u \in \tilde{C}^K(a,b) \quad (41)$$

$(\cdot\,)_K$ and $\|\cdot\|_K$ are the Sobolev inner product and norm, respectively, of order K. The completion of $\tilde{C}^K(a,b)$ with respect to the norm $\|\cdot\|_K$ produces a larger space denoted by $H^{(K)}(a,b)$, a Sobolev space.

Function of compact support

A function u defined on (a,b) has compact support on it if numbers α, β exist such that $a < \alpha < \beta < b$ and $u(x) = 0$ for $a < x < \alpha$ and $\beta < x < b$. $\mathring{H}^K(a,b)$ is the completion with respect to $\|\cdot\|_K$ of the subset of $C^K(a,b)$ with compact support on (a,b) and it is also a Sobolev space, a Hilbert space with respect to the Sobolev inner product and norm.

$$\mathring{H}^K(a,b) = \{v \in H^K(a,b);\ v^{(i)}(a) = v^{(i)}(b) = 0,\ i = 0,1,\ldots K-1\} \quad (42$$

The subspace of $C'(a,b)$ with compact support on (a,b) is not the same as $V = \{v \in C^2[a,b];\ v(a) = v(b) = 0\}$ but their completions with respect to the Sobolev norm $\|\cdot\|_1$ are the same. $\bar{V} = \mathring{H}^1(a,b)$. The problem $\mathcal{L}u = g$, $u(a) = u(b) = 0$ associated with

$$f(u) = \int_a^b \left[\tfrac{1}{2} p(x)(u')^2 + \tfrac{1}{2} q(x) u^2 - g(x) u\right] dx, \quad u \in V =$$

$\{v \in C^2[a,b];\ v(a) = v(b) = 0\}$ has the variational

formulation : find $u \in \overset{\circ}{H}{}^1(a,b)$ such that

$$a(u,\eta) = G(\eta) \quad \forall\, \eta \in \overset{\circ}{H}{}^1(a,b).$$

x x x x x

SOBOLEV SPACES (Several Space Dimensions)

Ω = a bounded open region of R^n

$L_2(\Omega)$ = the set of functions u defined on Ω and with the property that $\int_\Omega u^2\, d\Omega < \infty$. $L_2(\Omega)$ is a Hilbert space with respect to the inner product $(u,v) = \int_\Omega u v\, d\Omega$ and the norm $\|u\| = \sqrt{(u,u)}$.

$L_2(\Omega)$ and $L_2(\bar{\Omega})$ do not have any essential differen[ce] in all cases since the boundary $\partial\Omega$ has measure zero of physical interest. Sobolev spaces are traditionally defined on open regions. Hence $L_2(\Omega)$.

$u(x_1, \ldots, x_n)$; $\alpha_1, \alpha_2, \ldots, \alpha_n$ are nonnegative integers and $|\alpha| = \sum_{i=1}^{n} \alpha_i$.

$$D^{(\alpha)} = D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1}\, \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}}, \quad u^{(\alpha)} = D^\alpha u$$

(multi-index notations)

$$\tilde{C}^K(\Omega) = \left\{ v \in C^K(\Omega) ;\ \int_\Omega (v^{(\alpha)})^2\, d\Omega < \infty ;\ 0 \le |\alpha| \le K \right\}$$

$(u, v)_K = \sum'_{0 \le |\alpha| \le K} (u^{(\alpha)}, v^{(\alpha)})$ (summation is taken over all α such that $0 \le |\alpha| \le K$). $(u, v)_K$ is the Sobolev inner product on $\tilde{C}^K(\Omega)$ and the corresponding Sobolev norm is

$$\|u\|_K = \left(\sum_{0 \le |\alpha| \le K} \| u^{(\alpha)} \|^2 \right)^{1/2} \qquad (45)$$

The Sobolev space $H^K(\Omega)$ is the completion of $\tilde{C}^K(\Omega)$ with respect to the Sobolev norm. It $(H^K(\Omega))$ is a Hilbert space. with respect to the inner product $\|,\|_K$ and norm $\| \cdot \|_K$

A function u defined on a bounded open set $\Omega \subset R$ has <u>compact support</u> on Ω if there exists a closed set $\Omega_0 \subset \Omega$ such that $u = 0$ everywhere on $\Omega - \Omega_0$.

$\mathring{H}^K(\Omega)$ = completion with respect to norm $\| \cdot \|_K$ of the subset of $C^K(\Omega)$ with compact support on Ω. If $\partial\Omega =$ is smooth,

$$\mathring{H}^K(\Omega) = \left\{ v \in H^K(\Omega) ; \frac{\partial^i v}{\partial \nu^i} = 0 \text{ on } \Gamma, \ i = 0, 1, \ldots, K-1 \right\}$$

interpreting the normal derivative $\frac{\partial^i v}{\partial \nu^i}$ in the generalized sense.

The Sobolev space $H^K(\Omega)$ is the completion of $\tilde{C}^K(\Omega)$ with respect to the Sobolev norm. If $\{u_n\}$ is a sequence of functions $\subset \tilde{C}^K(\Omega)$ and has

the property that $\lim_{m,n\to\infty} \|u_m - u_n\|_K = 0$, we have

(since $\|u\|_K = \left(\sum_{0\le|\alpha|\le K}' \|u^{(\alpha)}\|^2 \right)^{1/2}$),

$\lim_{m,n\to\infty} \|u_m^{(\alpha)} - u_n^{(\alpha)}\| = 0$ for all α such that $0 \le |\alpha| \le K$

and for all such α the completeness of $L_2(\Omega)$ determines a unique function $u_\alpha \in L_2(\Omega)$ satisfying $\lim_{n\to\infty} \|u_\alpha - u_n^{(\alpha)}\| = 0$. For $|\alpha| = 0$, $u = u_\alpha$ is an element of $H^K(\Omega)$. For $0 < |\alpha| \le K$, each u_α is one of its generalized partial derivatives. If $u \in C^{|\alpha|}(\bar{\Omega})$ then $u_\alpha = u^{(\alpha)}$.

x x x x x

LAX-MILGRAM LEMMA

Let V be an arbitrary Hilbert space with the inner product $(\,,\,)_V$ and norm $\|\cdot\|_V = (u,u)_V^{1/2}$ for all $u \in V$. $a : V \times V \to R$ is a mapping which is bilinear, bounded and coercive.

$$a(\alpha u + \beta v, w) = \alpha\, a(u,w) + \beta\, a(v,w)$$
$$a(w, \alpha u + \beta v) = \alpha\, a(w,u) + \beta\, a(w,v)$$
$u, v, w \in V$, $\alpha, \beta \in R$ (bilinearity)

A constant β exists such that

$$|a(u,v)| \le \beta \|u\|_V \|v\|_V, \quad u, v \in V \quad \text{(boundedness)}$$

A constant $\rho > 0$ exists such that

$$a(u,u) \ge \rho \|u\|_V^2, \quad u \in V \quad \text{(coerciveness)}$$

$$a(u,v) = a(v,u), \quad u, v \in V. \quad \text{(Symmetry)}$$

$G: V \to R$ is a mapping such that

(1) $G(\alpha u + \beta v) = \alpha G(u) + \beta G(v)$, $u, v \in V$, $\alpha, \beta \in R$

(2) there exists a constant δ such that

$$\|G(u)\| \leq \delta \|u\|_V, \quad u \in V.$$

Then G is a bounded linear functional on V.

The Lemma asserts that if $a(u, v)$ is a bilinear, bounded and coercive mapping of $V \times V \to R$ and G is a bounded linear functional on V, there exists a unique element $\hat{u} \in V$ such that

$$a(\hat{u}, u) = G(u) \quad \forall \; u \in V. \tag{47}$$

If a is symmetric also, then $\hat{u}$ is the unique global minimizer of

$$f(u) = \frac{1}{2} a(u, u) - G(u), \quad u \in V.$$

THE RITZ METHOD

This is a standard method for the determination of an approximate solution to a boundary value problem. The problem has to be formulated so as to validate the application of the Lax-Milgram Lemma. The conditions for this are that the problem can be described in the form

$$a(u, v) = G(u) \quad \forall u \in V \qquad (48)$$

where V is an arbitrary Hilbert space with inner product $(\,,\,)_V$ and norm $\|\ \|_V$, $a(u,v)$ is a bounded and coercive bilinear form over V, G is a bounded linear functional on V. The solution to the boundary value problem is identical with the element $\hat{u} \in V$ and the existence of $\hat{u}$ is asserted by the Lax-Milgram Lemma. If $a(\cdot,\cdot)$ is a symmetric bilinear form on V, then the energy functional

$$f(u) = \tfrac{1}{2} a(u,u) - G(u), \quad u \in V \qquad (49)$$

and

$$\min_{u \in V} f(u) = f(\hat{u}) \qquad (50)$$

In the Ritz method of finding an approximate solution, applicable in the setting of the Lax-Milgram lemma, discretization is made by selecting an N-dimensional subspace of V, $(V_N \subset V)$. The subspace V_N is itself a Hilbert space and the conditions of the Lax-Milgram lemma are satisfied when V is

replaced by V_N.

Theorem. Let the conditions of the Lax-Milgram Lemma be satisfied and let V_N be an N-dimensional subspace of V. There exists a unique element $\hat{u}_N \in V_N$ such that

$$a(\hat{u}_N, u) = G(u) \quad \forall\, u \in V_N \tag{51}$$

If $a(\cdot,\cdot)$ is also symmetric, then

$$\min_{u \in V_N} f(u) = f(\hat{u}_N), \quad f \text{ being the energy}$$

functional $f(u) = \frac{1}{2} a(u,u) - G(u)$; $u \in V$.

Take $a(\cdot,\cdot)$ to be symmetric and let $\{\phi_i\}$ $(i = 1, 2, \cdots, N)$ be a basis of V_N. Every $u \in V_N$ has the form

$$u = \sum_{i=1}^{N} \alpha_i \phi_i \tag{52}$$

where $\alpha_1, \cdots, \alpha_N \in R$.

$$f(u) = \frac{1}{2} a(u,u) - G(u) = \frac{1}{2} \alpha^T K \alpha - \alpha^T G = F(\alpha) \tag{5.}$$

$\alpha = (\alpha_1, \cdots, \alpha_N)^T$, $\quad G = [G(\phi_1), G(\phi_2), \cdots G(\phi_N)]^T$

$K = (K_{ij})$, $\quad K_{ij} = a(\phi_j, \phi_i) = a(\phi_i, \phi_j)$

$\alpha^T K \alpha = a(u,u) \geq \rho \|u\|_V^2 \geq 0$ and ρ is independent of u (coercive property of $a(u,v)$). Thus when $a(\cdot,\cdot)$ is a symmetric bilinear form with the coercive property, the stiffness matrix K is positive definite.

From (53), $\min_{u \in V_N} f(u) = \min_{\alpha \in R^N} F(\alpha)$. F is a quadratic functional over R^N with a positive definite Hessian.

{ Hessian of $F(\alpha) = (f_{ij}) = \dfrac{\partial^2 F(\alpha)}{\partial \alpha_i \partial \alpha_j} = (K)$

and K is positive definite, as seen above }:

F is uniquely minimized by that vector α that satisfies the equation

$$K \alpha = G \qquad (54)$$

The minimizer $\hat{u}_N$ of f over V_N is given by $u = \sum_{i=1}^{N} \alpha_i \phi_i$ with the α_i's found from the linear system (54).

<u>Construction of $\hat{u}_N$</u>

(1) Choose a basis $\{\phi_i\}$ $(i = 1, \cdots, N)$ for V_N

(2) Construct the matrices K $(K_{ij} = a(\phi_j, \phi_i))$ and $G = \{G(\phi_1), \cdots, G(\phi_N)\}^T$

(3) Solve the linear system $K\alpha = G$ (an $N \times N$ linear set of equations).

(4) $\hat{u}_N = \sum_{i=1}^{N} \alpha_i \phi_i$

The minimizer $\hat{u}_N$ depends only ~~on~~ the subspace V_N but not the actual basis functions $\{\phi_i\}$ employed to span the subspace. While this observation is theoretically correct, the actual choice of the basis

functions has a great effect on the accuracy of the solution of the linear system $K\alpha = G$ (54).

Basis functions used in the Ritz method are of two types: (i) dimension-independent and (ii) dimension-dependent. In the first type, we may pass from N to $N+1$ dimensions by the addition of ϕ_{N+1} to the earlier choice $\{\phi_i\}$ $(i = 1, 2, \ldots, N)$. In type (2) the old set of N functions $\{\phi_i\}$ is replaced by a new set of $N+1$ functions $\{\phi_i\}$

In type (1) we must select the set $\{\phi_i\}$ $(i = 1, \ldots, N)$ from the sequence $\{\phi_i\}$ $(i = 1, \ldots, \infty) \subset V$ such that the former set of N functions is linearly independent for each of the values $N = 1, 2, \ldots$ lest the stiffness matrix $K = K^{(N)}$ should be singular. If $K^{(N)}$ is nearly singular (i.e. the spectral condition number of K^N i.e. $K(K^N)$ is large) the numerical solutions of (54) are contaminated by large rounding errors and iterative methods will have a low rate of convergence. For effective computation, apart from the linear independence of the functions $\{\phi_i\}$, we must ensure that the spectral condition number of the stiffness matrix $K^{(N)}$ is small enough. A natural choice of the basis functions (ϕ_i) $(i = 1, \ldots \infty)$ is the set of eigenfunctions of the operator $\mathcal{L}$ that

occurs in the boundary value problem $\mathcal{L} u = g$, $u(a) = u(b) = 0$. If the conditions of the Lax-Milgram lemma are satisfied and the bilinear form $a(\cdot,\cdot)$ is symmetric, we have the eigen values λ_i defined from the set

$$\mathcal{L}\, \phi_i = \lambda_i\, \phi_i \qquad (i = 1, 2, 3, \ldots) \tag{55}$$

and $0 < \lambda_1 \leq \lambda_2 \leq \cdots$, $\lim_{i \to \infty} \lambda_i = \infty$.

$$a(\phi_i, \phi_j) = \lambda_i\, \delta_{ij} \qquad (i, j = 1, 2, 3, \ldots)$$

$K^{(N)}$ is then the diagonal matrix $(\lambda_1, \ldots \lambda_N)$ and the solution of (54) is

$$\alpha_i = G(\phi_i) / \lambda_i \qquad (i = 1, 2, \ldots) \tag{56}$$

However all this needs prior determination of the eigen functions and this, in itself, is a major task.

Basis functions of type (2) are dimension-dependent have however, the advantage of being reasonably easy to work with in the finite element method. Also the stiffness matrix will be sparse and this is a great computational gain. The spectral condition number $K(K^N)$ is not uniformly bounded with respect to N for the finite element basis functions, but its rate of growth is modest and for second order boundary

value problems in two space dimensions $K(K(N))$ typically increases only linearly with N and the growth is even slower in the case of three dimensions.

ERROR ANALYSIS OF THE RITZ METHOD

$$f = \frac{1}{2} a(u, u) - G(u), \quad u \in V \tag{57}$$

$a(\cdot,\cdot)$ is symmetric and

$$\min_{u \in V_N} f(u) = f(\hat{u}_N) \tag{58}$$

$$f(u) = f(\hat{u}) + \frac{1}{2} a(\hat{u}-u, \hat{u}-u), \quad u \in V \tag{59}$$

$$a(\hat{u}-\hat{u}_N, \hat{u}-\hat{u}_N) = \min_{u \in V_N} a(\hat{u}-u, \hat{u}-u)$$

and in energy norm notation,

$$\|\hat{u}-\hat{u}_N\|_a = \min_{u \in V_N} \|\hat{u}-u\|_a \tag{60}$$

This result shows that the Ritz method minimizes the the error $\hat{u}-\hat{u}_N$ in the energy norm over the subspace V_N. Thus $\hat{u}_N$ is an optimal estimate in the energy norm.

From $a(\hat{u}, u) = G(u)$ for $u \in V$

$a(\hat{u}_N, u) = G(u)$ for $u \in V_N$.

we see that $a(\hat{u}-\hat{u}_N, u) = 0$ for $u \in V_N$.

Geometrically, this means that $\hat{u}_N$ is the orthogonal projection ~~of $\hat{u}$ onto V_N~~ with respect to the energy inner product of $\hat{u}$ onto V_N. If $\{\phi_i\}$ is a set

of (dimension-independent) basis functions in V and are orthonormal with respect to the energy inner product, i.e. $a(\phi_i, \phi_j) = \delta_{ij}$ and the set is complete in V, we have

$$\hat{u} = \sum_{i=1}^{\infty} b_i \phi_i \quad , \quad b_i = a(\hat{u}, \phi_i)$$

If $V_N = \text{SPAN}\{\phi_1, \ldots, \phi_N\}$, $N = 1, 2, \ldots$ the partial sum $\sum_{i=1}^{N} b_i \phi_i$ is orthogonal to the projection of $\hat{u}$ onto V_N with respect to the energy inner product. $\hat{u}_N = \sum_{i=1}^{N} b_i \phi_i$

$$\hat{u} - \hat{u}_N = \sum_{i=N+1}^{\infty} b_i \phi_i$$

$$\|\hat{u} - \hat{u}_N\|_{\mathcal{L}} = \left(\sum_{i=N+1}^{\infty} b_i^2 \right)^{1/2} \qquad (61)$$

Since $a(\cdot,\cdot)$ is a bounded and coercive bilinear form, there exist positive constants ρ and β such that

$$\sqrt{\rho}\, \|u\|_V \leq \sqrt{a(u,u)} \leq \beta \|u\|_V \quad \forall\, u \in V \qquad (62)$$

From the result

$$a(\hat{u} - \hat{u}_N, \hat{u} - \hat{u}_N) = \min_{u \in V_N} a(\hat{u} - u, \hat{u} - u) \qquad (63)$$

we see that

$$\|\hat{u} - \hat{u}_N\|_V \leq \rho^{-1/2} \{a(\hat{u} - \hat{u}_N, \hat{u} - \hat{u}_N)\}^{1/2}$$

$$\leq (\rho^{-1}\beta)^{1/2} \min_{u \in V_N} \|\hat{u} - u\|_V \qquad (64)$$

This is a <u>quasi-optimal</u> estimate.

THE GALERKIN METHOD

Let $a(u, v)$ be a bilinear form but not symmetric necessarily. It is bounded and satisfies the coercive property. Let V_N be an N-dimensional subspace of V. By Lax-Milgram Lemma, we see that there exists a unique element $\hat{u}_N \in V_N$ such that

$$a(\hat{u}_N, u) = G(u) \quad \forall u \in V_N \tag{65}$$

$$\hat{u}_N = \sum_{i=1}^{N} \alpha_i \phi_i \tag{66}$$

if $\{\phi_i\}_{i=1,2,\ldots N}$ is a basis for V_N. From (65), it follows that

$$\sum_{j=1}^{N} \alpha_j \, a(\phi_j, u) = G(u) \quad \forall u \in V_N$$

and hence

$$\sum \alpha_j \, a(\phi_j, \phi_i) = G(\phi_i) \quad i = 1, 2, \ldots, N. \tag{67}$$

$$K\alpha = G$$

$$\text{with} \quad K_{ij} = a(\phi_j, \phi_i) \tag{68}$$

Linear independence of the basis functions $\phi_i \; (i=1,\ldots,N)$ and the coercive property of $a(\cdot,\cdot)$ imply the nonsingular nature of the matrix K. $\bar{\alpha}$ is then determined from the linear system (67). However, the stiffness matrix K is nonsymmetric and the numerical computation is more costly than in the method of Ritz.

Error estimate in the norm $\|\cdot\|_V$.

The assumptions of the Lax-Milgram Lemma are satisfied, $V_N \subset V$ is an N-dimensional subspace of V; $\hat{u}$ and $\hat{u}_N$ are the elements identified in the L-M lemma and the theorem regarding the existence of $\hat{u}_N$.

$$a(\hat{u}_N, u) = G(u) \quad \forall u \in V_N \tag{69}$$

$$a(\hat{u}, u) = G(u) \quad \forall u \in V \tag{70}$$

$$\therefore \quad a(\hat{u} - \hat{u}_N, u) = 0 \quad \forall u \in V_N \tag{71}$$

$$a(\hat{u} - \hat{u}_N, \hat{u} - \hat{u}_N) = a(\hat{u} - \hat{u}_N, \hat{u} - u + u - \hat{u}_N)$$

$$= a(\hat{u} - \hat{u}_N, \hat{u} - u) + a(\hat{u} - \hat{u}_N, u - \hat{u}_N)$$

$$= a(\hat{u} - \hat{u}_N, \hat{u} - u) + 0 \tag{72}$$

since $u - \hat{u}_N \subset V_N \quad \forall u \subset V_N$ and

$$a(\hat{u} - \hat{u}_N, u) = 0 \quad \forall u \in V_N. \tag{73}$$

The boundedness and coercive property lead to

$$\rho \|\hat{u} - \hat{u}_N\|_V^2 \leq a(\hat{u} - \hat{u}_N, \hat{u} - \hat{u}_N) = a(\hat{u} - \hat{u}_N, \hat{u} - u)$$

$$\leq \beta \|\hat{u} - \hat{u}_N\|_V \|\hat{u} - u\|_V \quad \forall u \in V_N.$$

It follows that $\|\hat{u} - \hat{u}_N\|_V \leq \rho^{-1} \beta \|\hat{u} - u\|_V \quad \forall u \in V$

Thus
$$\|\hat{u} - \hat{u}_N\|_V \leq \rho^{-1} \beta \min_{u \in V_N} \|\hat{u} - u\|_V \tag{74}$$

(A quasi-optimal error estimate in the norm $\|\cdot\|_V$.

Finite Element Basis Functions

L : number of elements in a finite element mesh.

M : number of nodes in a finite element mesh

T : number of nodes in a single element

e_l $(l = 1, 2, \dots, L)$: the elements

$N_i = (x_i, y_i) \mid (x_i, y_i, z_i)$ the nodes

i_r^l $(r = 1, 2, \dots, T)$: node numbers in the element e_l

$\phi_i(x, y)$ $(i = 1, \dots, M)$ Global basis functions defined on $\bar{\Omega}$

$\phi_r^{(l)}(x, y)$ $(r = 1, 2, \dots, T)$ Local basis functions defined on e_l

$\tilde{e}$ standard element

$\tilde{\phi}_r(\tilde{x}, \tilde{y})$ $(r = 1, 2, \dots, T)$ standard basis functions defined on $\tilde{e}$.

$\Omega \subset R^2$ is an open, bounded, simply connected domain and is the domain of definition of the boundary value problem. $\bar{\Omega}$ = closure of Ω and $\bar{\Omega}$ is divided into the closed subregions or elements $\{e_l\}$, $l = 1, \dots, L$ e.g. e_l are polygons if $\bar{\Omega}$ is itself a polygon. If $\bar{\Omega}$ is not a polygon, curved elements are unavoidable.

In $\bar{\Omega}$ we select nodes $N_i = (x_i, y_i)$ $(i = 1, \dots, M)$ which are vertices of elements or edge points of elements or interior points of elements. A vertex

node can belong to several elements, an edge node can belong to at most two elements and an internal node can belong to only one element. The elements and nodes together constitute a <u>finite element mesh</u>. A simplifying assumption is that all elements in a mesh are the same type of polygons with the same number of vertices, edges and internal nodes.

The generation of a suitable finite element mesh over a given domain is not a trivial task and a set of general rules for generating a finite element mesh is indicated below.

(1) Very large and very small element angles are to be avoided; (2) where the solution of the problem or its derivatives vary rapidly, there is to be a dense selection of elements; (3) a fine mesh or many nodes per element are needed if the aim is high accuracy. However, many nodes per element yield high accuracy, only if the solution is sufficiently smooth.

Global node numbers of nodes in an arbitrary element e_l are $i_r^{(l)}$ $(r = 1, 2, \ldots, T)$; r is the local node number.

Finite Element Basis Functions

Suppose that for a given finite element mesh there is associated with each node $N_i = (x_i, y_i)$ a function $\phi_i = \phi_i(x,y)$ defined on Ω with the following properties

(1) $\phi_i(x,y) = \sum_{s=1}^{T} c_{il}^{(s)} x^{p_s} y^{q_s} \quad (x,y) \in e_l$. (1)

where T = the total number of nodes in the element e_l .

(2) $\phi_i(N_j) = \delta_{ij} \quad (i, j = 1, 2, \ldots, M)$ (2)

(3) ϕ_i is uniquely determined on every element edge by its values at the nodes belonging to that edge .

$$V_M = \text{SPAN}\{\phi_1, \phi_2, \ldots, \phi_M\} \quad (3)$$

is a finite element space and $\phi_1, \phi_2, \ldots, \phi_M$ form a finite element basis of functions. The functions $\phi_1, \ldots \phi_M$ are linearly independent and V_M has dimension M. The coefficients $c_{il}^{(s)}$ in (1) are determined from the linear system (2), i.e

$$\sum_{i=1}^{T} c_{il}^{(s)} \, x_{i_r^{(l)}}^{p_s} \, y_{i_r^{(l)}}^{q_s} = \begin{cases} 1 & \text{if } i_r^{(l)} = i \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$(i = 1, 2, \ldots, T)$$

$i_r^{(l)}$ = the global node number, explained in the figure below for a finite element mesh with 6 triangular elements, total number of 40 nodes and 10 nodes in each element

($L = 6$, $M = 40$, $T = 10$).

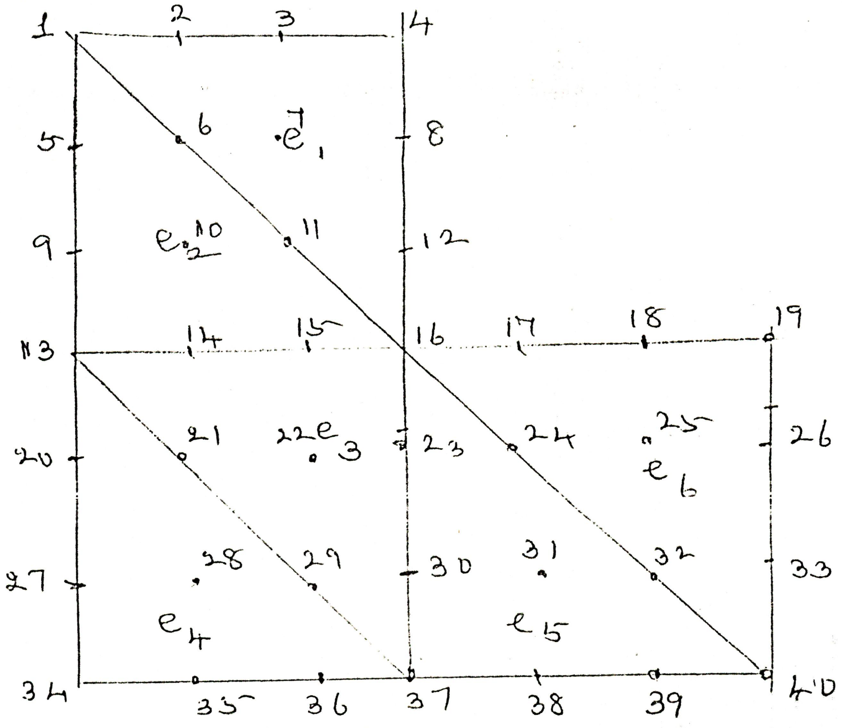
The system (4) is described in the form

(4)

$$S_L\, c_{iL} = d_{iL} \qquad (5)$$

$$S_L = \begin{pmatrix} x_{i_1^L}^{p_1} y_{i_1^L}^{q_1} & x_{i_1^L}^{p_2} y_{i_1^L}^{q_2} & \cdots & x_{i_1^L}^{p_T} y_{i_1^L}^{q_T} \\ x_{i_2^L}^{p_1} y_{i_2^L}^{q_1} & x_{i_2^L}^{p_2} y_{i_2^L}^{q_2} & \cdots & x_{i_2^L}^{p_T} y_{i_2^L}^{q_T} \\ \vdots & & & \\ x_{i_T^L}^{p_1} y_{i_T^L}^{q_1} & x_{i_T^L}^{p_2} y_{i_T^L}^{q_2} & \cdots & x_{i_T^L}^{p_T} y_{i_T^L}^{q_T} \end{pmatrix} \qquad (6)$$

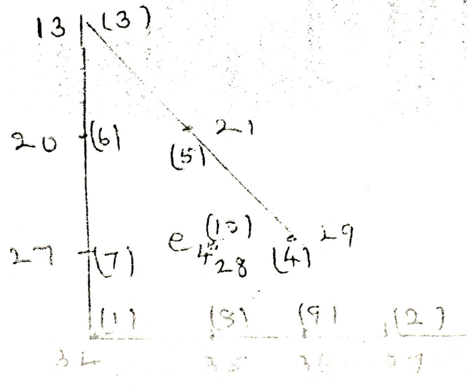
S_L is nonsingular (except in special 'degenerate' cases) and ϕ_L is well defined on e_L.



Elements $e_1, e_2, e_3, e_4, e_5, e_6$ $(L = 6)$

Nodes (global) $1, 2, \ldots, 40$ $(N = 40)$

In each element e_L, there are 10 nodes $(T = 10)$



$i_1^{(4)}$ = first local node in element e_4 = node 34.

$$\begin{cases} i_1^{(4)} = 34, \quad i_2^{(4)} = 37, \quad i_3^{(4)} = 13 \\ i_4^{(4)} = 29, \quad i_5^{(4)} = 21 \end{cases}$$

ϕ_i is well defined on e_l and is continuous at element edges. Since all polynomials are continuous, we have the property

(4) $\phi_i \in C(\bar{\Omega})$ (7)

$\phi_i = 0$ in every element e_l which does not contain the node N_i

(5) ϕ_i assumes nonzero values only in those elements to which the node N_i belongs. ϕ_i vanishes outside its support.

The union of elements to which N_i belongs is the support of the function ϕ_i. If N_i is not on $\Gamma = \partial\Omega$, $\phi_i = 0$ on the boundary of its support. If N_i is on $\Gamma = \partial\Omega$, $\phi_i = 0$ on that part of the boundary of its support that lies in Ω.

$$u(x, y) = \sum_{i=1}^{M} c_i \, \phi_i(x, y) \, , \quad (x, y) \in \bar{\Omega} \tag{8}$$

u is a continuous, piecewise polynomial over $\bar{\Omega}$ and belongs to $V_M = \text{SPAN}\{\phi_1, \phi_2, \ldots, \phi_M\}$.

Let
$$u_I(x, y) = \sum_{i=1}^{M} u(N_i) \, \phi_i(x, y) \quad (x, y) \in \bar{\Omega} \tag{9}$$

Then $u_I(N_i) = u(N_i) \quad (i = 1, 2, \ldots, M)$ and the function $u_I(x, y)$ in (9) interpolates u (collocates with u at the nodes). We may call u_I the V_M interpolant of u and

$$u_I(x, y) = u(x, y) \quad \forall (x, y) \in \bar{\Omega} \tag{10}$$

is equivalent to saying that $u(x, y) \in V_M$.

Let $P_K(\bar{\Omega})$ = the space of all polynomials of degree $\leq K$, defined on $\bar{\Omega}$; these are global and not piecewise, polynomials. K is the largest integer such that every term of the form $x^{K_1} y^{K_2}$ occurs with $K_1 + K_2 \leq K$. The terms necessary for a complete polynomial of degree K are

$(K = 0 : 1)$; $(K = 1 : x, y)$; $(K = 2; x^2, xy, y^2)$; (11)

$(K = 3; x^3, x^2y, xy^2, y^3)$; . . . . .

$u \in P_K(\bar{\Omega})$ implies that $u_1(x,y) = u(x,y) \ \forall (x,y) \in \bar{\Omega}$

This means that all polynomials of degree $\leq K$ are interpolated exactly in V_M if K and m are the greatest integers for which

$$P_K(\bar{\Omega}) \subseteq V_M \subset C^{m-1}(\bar{\Omega}). \quad (12)$$

Since $\phi_i \in C(\bar{\Omega})$ and $u(x,y) = \sum_{i=1}^{M} c_i \phi_i(x,y)$, $V_M = \text{SPAN}\{\phi_1, \phi_2, \ldots, \phi_M\}$, it is clear that $m \geq 1$ in (12). The normal derivative of a basis function at element edges in its support is discontinuous, for the basis functions under discussion and then $V_M \notin C^1(\bar{\Omega})$

Thus $m = 1$. The function $u_1(x,y)$ interpolates the values of u at the nodes but derivatives of u_1 do not interpolate the derivatives of u. This is Lagrange interpolation. If in addition; derivatives of u_1 are also to interpolate derivatives of u, we have Hermite interpolation. Finite element basis functions with $m = 1$ are appropriate for boundary value problems in which the space V of the

of the Lax-Milgram Lemma is $H^1(\Omega)$ or a subspace of $H^1(\Omega)$. For fourth order boundary value problems we need $m = 2$.

<u>Local Basis Functions</u>. These are defined over the elements e_l

$$\phi_r^{(l)}(x,y) = \phi_{i_r^{(l)}}(x,y), \quad (x,y) \in e_l, \quad r = 1,2,\dots,T \quad (13)$$

$V_T^{(l)}$ is the space of all polynomials of the type (13) defined on e_l and

$$V_T^{(l)} = \text{SPAN}\{\phi_1^{(l)}, \phi_2^{(l)}, \dots, \phi_T^{(l)}\} \quad (14)$$

Obviously, $\phi_r^{(l)}(N_{i_s^{(l)}}) = \delta_{rs}, \quad (r,s = 1,\dots T)$

and
$$u_1(x,y) = \sum_{i=1}^{M} u(N_i)\,\phi_i(x,y)$$
$$= \sum_{r=1}^{T} u(N_{i_s^{(l)}})\,\phi_r^{(l)}(x,y), \quad (x,y) \in e_l \quad (15)$$

Standard (reference) element $\tilde{e}$ in the $(\tilde{x}+\tilde{y})$ plane.

Choose the reference element $\tilde{e}$ with the local basis functions $\tilde{\phi}_1(\tilde{x},\tilde{y}), \tilde{\phi}_2(\tilde{x},\tilde{y}), \dots \tilde{\phi}_T(\tilde{x},\tilde{y})$ and find for every element e_l a non singular affine transformation

$$x = x(\tilde{x},\tilde{y}) = f_{11}\tilde{x} + f_{12}\tilde{y} + b_1, \quad (\tilde{x},\tilde{y}) \in \tilde{e} \quad (16)$$
$$y = y(\tilde{x},\tilde{y}) = f_{21}\tilde{x} + f_{22}\tilde{y} + b_2$$

The inverse transformation is

$$\tilde{x} = \tilde{x}(x,y), \quad \tilde{y} = \tilde{y}(x,y) \quad (17)$$

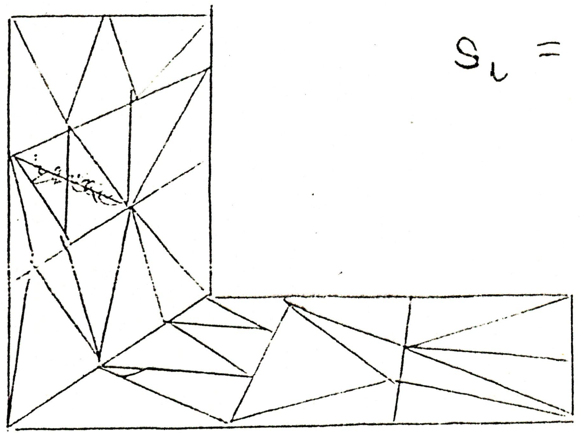
(16) maps $\tilde{e}$ onto e_l, mapping nodes of $\tilde{e}$ onto nodes of e_l and

$$\phi_r^{(l)}(x,y) = \tilde{\phi}_r^{(l)}(\tilde{x}(x,y), \tilde{y}(x,y)) \quad (18)$$

The local basis functions $\phi_{r}^{(l)}(x,y)$ $(l = 1, 2, \ldots, L)$ can be reduced by variable transformations to the single function $\tilde{\phi}_{r}$. The global basis functions $\phi_1, \phi_2, \ldots, \phi_M$ are composites of the local basis functions and <u>all operations on the global basis functions can be reduced to operations on $\tilde{\phi}_1, \tilde{\phi}_2, \ldots \tilde{\phi}_T$</u>. These are the standard (reference) local basis functions.

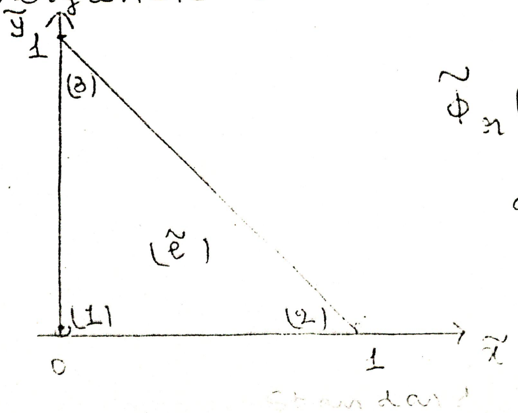
<u>Examples</u> 1. Piecewise Linear Basis Functions on Triangular Elements. The elements are triangles with nodes at the vertices and

$$\phi_i(x,y) = c_{il}^{(1)} + c_{il}^{(2)} x + c_{il}^{(3)} y, \quad (x,y) \in e_l \qquad (19)$$



$$S_l = \begin{pmatrix} 1 & x_{i_1^{(l)}} & y_{i_1^{(l)}} \\ 1 & x_{i_2^{(l)}} & y_{i_2^{(l)}} \\ 1 & x_{i_3^{(l)}} & y_{i_3^{(l)}} \end{pmatrix} \qquad (20)$$

S_l is nonsingular unless the triangular element e_l degenerates to a straight line. An element edge is described by the equation $ax + by + c = 0$. ϕ_i is determined everywhere on the edge by its values at the two nodes.



$$\tilde{\phi}_{r}(x,y) = \tilde{c}_{r}^{(1)} + \tilde{c}_{r}^{(2)} \tilde{x} + \tilde{c}_{r}^{(3)} \tilde{y} \qquad (21)$$

and the coefficients are such that $\tilde{\phi}_{r}$ has the value unity

Standard element

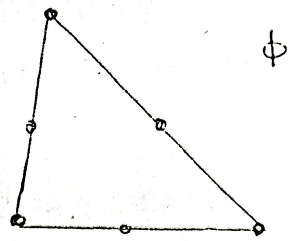
at node r and zero at the two other nodes.

$$\tilde{\phi}_1(\tilde{x},\tilde{y}) = 1-\tilde{x}-\tilde{y}, \quad \tilde{\phi}_2(\tilde{x},\tilde{y}) = \tilde{x}, \quad \tilde{\phi}_3(\tilde{x},\tilde{y}) = \tilde{y} \qquad (22)$$

The parameter $K=1$ and $m=1$.

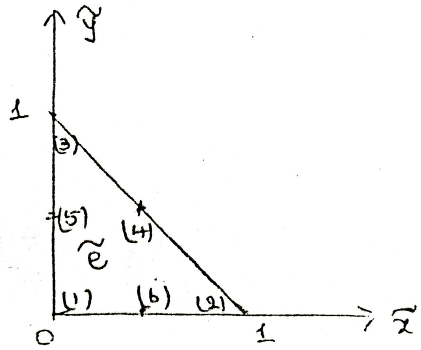
(2) Piecewise quadratic functions on Triangular Elements.

Elements are triangles with nodes at the vertices and middle points of edges.



$$\phi_i(x,y) = c_{il}^{(1)} + c_{il}^{(2)} x + c_{il}^{(3)} y + c_{il}^{(4)} x^2 + c_{il}^{(5)} xy + c_{il}^{(6)} y^2 \qquad (23)$$

$$(x,y) \in e_l$$



Standard element $\tilde{e}$,

The standard local basis functions are

$$\tilde{\phi}_1(\tilde{x},\tilde{y}) = \tilde{\lambda}_1(2\tilde{\lambda}_1-1), \quad \tilde{\phi}_2(\tilde{x},\tilde{y}) = \tilde{\lambda}_2(2\tilde{\lambda}_2-1),$$

$$\tilde{\phi}_3(\tilde{x},\tilde{y}) = \tilde{\lambda}_3(2\tilde{\lambda}_3-1), \quad \tilde{\phi}_4(\tilde{x},\tilde{y}) = 4\tilde{\lambda}_2\tilde{\lambda}_3$$

$$\tilde{\phi}_5(\tilde{x},\tilde{y}) = 4\tilde{\lambda}_3\tilde{\lambda}_1, \quad \tilde{\phi}_6(\tilde{x},\tilde{y}) = 4\tilde{\lambda}_1\tilde{\lambda}_2 \qquad (24)$$

$\tilde{\lambda}_1 = 1-\tilde{x}-\tilde{y}$, $\tilde{\lambda}_2 = \tilde{x}$, $\tilde{\lambda}_3 = \tilde{y}$ are the standard local basis functions of the previous example.

Ordering of the nodes is based on the following

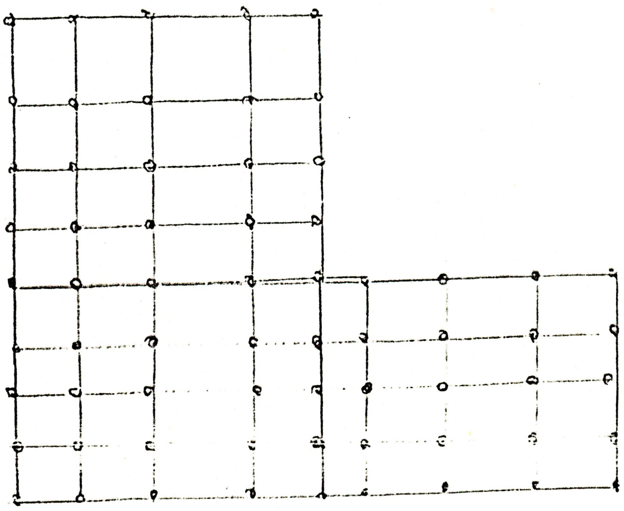
convention: The vertex nodes are ordered first in the counterclockwise sense and the nodes 4, 5, 6 are at the middle points of the edges opposite to the vertices 1, 2, 3 respectively.

For the basis functions of this example $K = 2$ and $m = 1$.

(3) Rectangular Elements; piecewise bilinear basis functions.

The elements are <u>axiparallel</u> rectangles with nodes at the vertices and

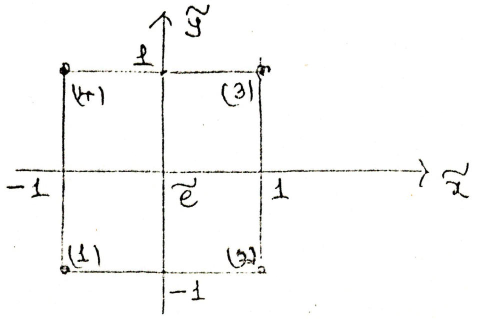
$$\phi_i(x, y) = c^{(1)}_{il} + c^{(2)}_{il}\, x + c^{(3)}_{il}\, y + c^{(4)}_{il}\, xy \qquad (25)$$



Typical mesh with rectangular elements.

On any element edge ϕ_i reduces to a linear function of x or y and is uniquely determined by its values at the two vertex nodes.

Element $\tilde{e}$ (standard element)



$$\left.\begin{aligned}
\tilde{\phi}_1(\tilde{x}, \tilde{y}) &= \tfrac{1}{4}(1-\tilde{x})(1-\tilde{y}) \\
\tilde{\phi}_2(\tilde{x}, \tilde{y}) &= \tfrac{1}{4}(1+\tilde{x})(1-\tilde{y}) \\
\tilde{\phi}_3(\tilde{x}, \tilde{y}) &= \tfrac{1}{4}(1+\tilde{x})(1+\tilde{y}) \\
\tilde{\phi}_4(\tilde{x}, \tilde{y}) &= \tfrac{1}{4}(1-\tilde{x})(1+\tilde{y})
\end{aligned}\right\} \qquad (26)$$

The support of ϕ_i is always comprising four elements unless N_i belongs to $\Gamma = \partial\Omega$.

Implementation of the Ritz – Galerkin method with finite element basis functions.

Consider the boundary value problem

$$-\{(p u_x)_x + (p u_y)_y\} + q u = g \quad \text{in } \Omega \tag{27}$$

$$u = \gamma \text{ on } \Gamma_0 , \quad u_\nu + \sigma u = \xi \text{ on } \Gamma_1 \tag{28}$$

Γ_0 = part of $\Gamma = \partial\Omega$ where an essential boundary condition is given. M_0 = number of nodes in Γ_0.

J_0 = set of node numbers of nodes on Γ_0

Γ_1 = part of $\Gamma = \partial\Omega$ where a natural boundary condition is given.

L_1 = number of element edges on Γ_1

N = number of nodes in $\Omega \cup \Gamma_1$

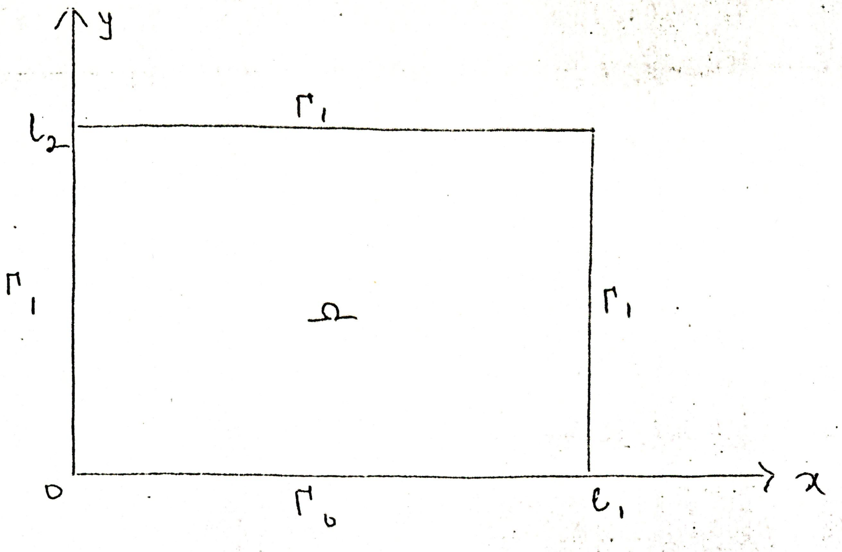
J = set of node numbers of nodes in $\Omega \cup \Gamma_1$

Q = number of integration points in the standard element $\tilde{e}$

$(\tilde{x}^{(m)}, \tilde{y}^{(m)})$ $m = 1, 2, \ldots, Q$ integration points

w_m, $(m = 1, 2, \ldots Q)$ integration weights.

The domain of the problem is Ω shown in the picture below. The finite element method has great flexibility in respect of the geometry of the domain



Domain of the problem.

and if Γ is curved, it cannot be represented exactly by the boundary of a finite element mesh.

Let $\quad H^1_\gamma(\Omega) = \{ v \in H^1(\Omega);\ v = \gamma \text{ on } \Gamma_0 \}$ (29)

$H^1_0(\Omega) = \{ v \in H^1(\Omega);\ v = 0 \text{ on } \Gamma_0 \}$ (30)

The variational formulation of the problem is:

to find a function $\hat{u} \in H^1_\gamma(\Omega)$ that satisfies

$$a(\hat{u}, v) = G(v) \quad \forall\, v \in H^1_0(\Omega) \tag{31}$$

with

$$a(u,v) = \iint_\Omega \{ p\, u_x v_x + p\, u_y v_y + q\, u v \}\, dx\, dy + \int_{\Gamma_1} p \sigma u v\, ds$$

$$= a(u,v)_\Omega + a(u,v)_{\Gamma_1} \tag{32}$$

$$G(v) = \iint_\Omega g\, v\, dx\, dy + \int_{\Gamma_1} p \xi v\, ds$$

$$= G(v)_\Omega + G(v)_{\Gamma_1} \tag{33}$$

$\hat{u}$ minimizes the functional

$$f(u) = \tfrac{1}{2} a(u,u) - G(u), \quad u \in H^1_\gamma(\Omega).$$

Let u^* be any function in $H^1_\gamma(\Omega)$. Then

$$H^1_\gamma(\Omega) = \{ u \in H^1(\Omega) ;\ u = v + u^*,\ v \in H^1_0(\Omega) \} \quad (34)$$

and the problem (31) is equivalent to finding $\hat{v} \in H^1_0(\Omega)$ such that

$$a(\hat{v} + u^*, v) = G(v) \quad \forall\, v \in H^1_0(\Omega) \quad (35)$$

V_N = an N-dimensional subspace of $H^1_0(\Omega)$ with basis functions $\phi_1, \ldots, \phi_N$. The Ritz-Galerkin equations for the problem are

$$K \alpha = G^* \quad (36)$$

$$K = K_{ij}, \quad K_{ij} = a(\phi_j, \phi_i) \quad i, j = 1, 2, \ldots, N$$

$$G_i^* = G(\phi_i) - a(u^*, \phi_i) \quad i = 1, 2, \ldots, N$$

The Ritz-Galerkin approximations to $\hat{v}$, $\hat{u}$ are

$$\hat{v}_N = \sum_{i=1}^{N} \alpha_i \phi_i \quad \text{and} \quad \hat{u}_N = \hat{v}_N + u^* \quad (37)$$

Finite Element Basis Functions

A mesh of L elements and M nodes on $\bar{\Omega}$ and finite element basis functions $\phi_1, \phi_2, \cdots, \phi_M$ with the properties (1) to (7) below.

(1) ϕ_i has the polynomial form

$$\phi_i(x,y) = \sum_{s=1}^{T} c_{il}^{(s)} x^{p_s} y^{q_s} \qquad (x,y) \in e_l \tag{38}$$

and the powers p_s, q_s $(s = 1, 2, \cdots, T)$ are independent of i and l.

(2) $\phi_i(N_j) = \delta_{ij} \qquad (i, j = 1, 2, \cdots, M).$ (39)

(3) ϕ_i is uniquely determined on every edge by its values at the nodes belonging to that edge.

(4) $\phi_i \in C(\bar{\Omega})$ (40)

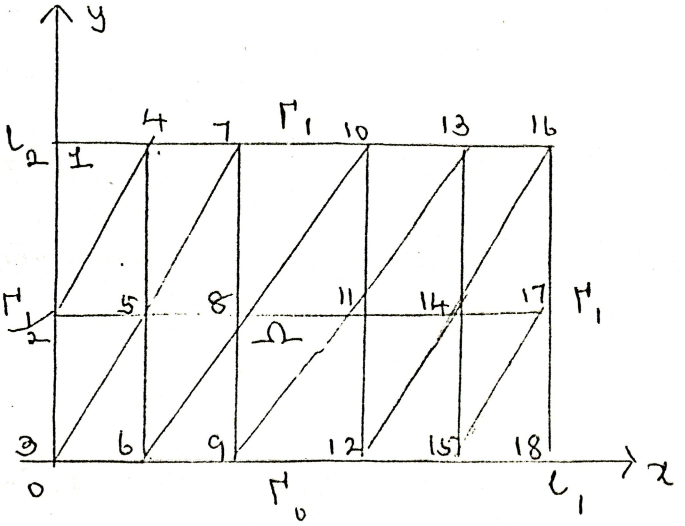
(5) ϕ_i assumes nonzero values only on those elements to which N_i belongs.

(6) If N_i is not on $\Gamma = \partial\Omega$, ϕ_i vanishes on the boundary of its support (Support of ϕ_i is the union of those elements to which N_i belongs). If N_i is on Γ, ϕ_i vanishes on that part of the boundary of its support that lies in Ω.

(7) It is possible to choose a standard or reference element $\tilde{e}$ with the local basis functions $\tilde{\phi}_1(\tilde{x}, \tilde{y})$, $\tilde{\phi}_2(\tilde{x}, \tilde{y})$, $\cdots$, $\tilde{\phi}_T(\tilde{x}, \tilde{y})$ in polynomial form as

in (1) above

$$\left.\begin{aligned} -(p u_x)_x - (p u_y)_y + q u &= g \quad \text{in } \Omega \\ u &= r \quad \text{on } \Gamma_0 \\ u_\nu + \sigma u &= \xi \quad \text{on } \Gamma_1 \end{aligned}\right\} \qquad (41)$$



M_0 : Number of nodes on Γ_0 (6)

J_0 : Set of node numbers of nodes on Γ_0 $\{3, 6, 9, 12, 15, 18\}$

N : number of nodes in $\Omega \cup \Gamma_1$ $(= M - M_0)$ $(= 12)$

J : set of node numbers of nodes in $\Omega \cup \Gamma_1$

$= \{1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17\}$

J_0 : nodes at which $\hat{u}$ is known

J : nodes at which $\hat{u}$ is unknown.

Place nodes at the end points of $\bar{\Gamma}_0$ and assign them to Γ_0. Nodes 3, 18 are assigned to Γ_0.

We have to find $u^* \in H^1_r(\Omega)$. Take

$$u^*_N(x, y) = \sum_{i \in J_0} r(N_i)\, \phi_i(x, y) \qquad (x, y) \in \bar{\Omega}$$

$u^*_N \in H^1(\Omega)$ and $u^*_N(N_i) = r(N_i)$ for $i \in J_0$

Thus u^*_N interpolates r at all the boundary nodes on Γ_0 but may not interpolate r between the nodes.

Thus $u_N^* \in H^1(\Omega)$ but $\notin H_N^1(\Omega)$. If the mesh is fine enough, u_N^* will be a good approximation to some function in $H_N^1(\Omega)$ and we may replace u^* in $K\alpha = G^*$ by u_N^* (This has been called a "variational crime" by Strang and Fix [An analysis of the Finite Element Method" (1973 - Prentice Hall].)

To determine a suitable subspace $V_N \in H_0^1(\Omega)$.

$$V_N = \text{SPAN}\{\phi_i, \quad i \in J\} \tag{42}$$

There are N basis functions and each of them is such that $\phi_i \in H^1(\Omega)$ and $\phi_i = 0$ on Γ_0. V_N is a sub-space of $H_0^1(\Omega)$ as required. The Ritz-Galerkin equations become

$$K_{ij} = a(\phi_j, \phi_i) \qquad i, j \in J \tag{43}$$

$$G_i^* = G(\phi_i) - \sum_{j \in J_0} \gamma(N_j)\, a(\phi_j, \phi_i) \tag{44}$$

These data are to be computed and for this task we need two programming arrays with dimensions $M \times M$ and $M \times 1$ respectively for the stiffness matrix K_{ij} ~~$= a(\phi_j, \phi_i) \quad i, j \in J$~~ and the column G_i^*. The size of the arrays is determined by M and not N to make the computations as much independent of the boundary conditions as possible.

$$(1)\ K[i, j] = a(\phi_j, \phi_i)_\Omega \qquad i, j = 1, 2, \ldots, M \tag{45}$$

(2) $G[i] := G(\phi_i)_{\Omega}$, $\quad i = 1, 2, \dots, M$ (46)

(3) $K[i,j] := K[i,j] + a(\phi_j, \phi_i)_{\Gamma_1}$ $\quad i,j = 1, \dots M$ (47)

(4) $G[i] := G[i] + G(\phi_i)_{\Gamma_1}$ $\quad i = 1, \dots M$ (48)

(5) $\forall\, j \in J_0$

$G[i] := G[i] - \gamma(N_j)\, K[i,j]$ $\quad \forall\, i \in J$

$K[i,j] := 0, \quad K[j,i] := 0$ $\quad \forall\, i \in J$

$K[j,i] := 0 \quad \forall\, i \in J_0, \quad i \neq j$

$K[j,j] := 1; \quad G[j] := \gamma(N_j)$ (49)

In the first two steps boundary conditions are not involved but they are present in the next two steps and in the step (5) the essential boundary conditions are taken into account.

For every $i \in J_0$ the i th row and column contains zeros at the end of the computation of the array K except in the diagonals where $K(i,i) = 1$. and $G(i) = \gamma(N_i)$. If all such rows and columns are removed, the arrays K and G assume the dimensions $N \times N$ and $N \times 1$ respectively and contain precisely the data $K(i,j) = K_{ij} = a(\phi_j, \phi_i)$,

$$G_i^* = G(\phi_i) - \sum_{j \in J_0} \gamma(N_j)\, a(\phi_j, \phi_i) \quad \text{for } i, j \in J. \tag{50}$$

We may also append the $M - N$ equations

$$u(N_i) = \gamma(N_i) \quad \forall\, i \in J_0 \tag{51}$$

to the original N th order Ritz - Galerkin system.

After the nodes of $\Omega \cup \Gamma_1$ are ordered before those of Γ_0 we see that the contents of the matrix K assume the structured block form

$$
\begin{array}{c}
N \left\{ \vphantom{\begin{matrix} a \\ b \\ c \end{matrix}} \right. \\
M-N \left\{ \vphantom{\begin{matrix} a \end{matrix}} \right.
\end{array}
\left[
\begin{array}{c|c}
K & O_{N,\,M-N} \\
\hline
O_{M-N,\,N} & I_{M-N,\,M-N}
\end{array}
\right]
\begin{array}{c}
\vphantom{\begin{matrix} a \\ b \\ c \end{matrix}} \\
\left. \vphantom{\begin{matrix} a \end{matrix}} \right\} M-N
\end{array}
$$

$$\underbrace{\qquad\quad}_{N}\qquad\underbrace{\qquad}_{M-N}$$

When the mesh is fine, N/M is close to unity and the extra equations to be solved are simple so that the above ordering need not be done with a view to delete or detach the extra equations M − N from the set of N core equations controlled by the block K in the above diagram.

If the equation solver implements the simple conjugate gradient method, the node ordering is not that important. If however, it implements the conjugate gradient method with preconditioning or implements a direct method, the node ordering has an important effect on the computational labour and possibly data storage as well.

The stiffness matrix K of order M × M that does not take the boundary conditions into account is the <u>primitive</u> matrix.

The symmetry of $a(\cdot,\cdot)$ can be taken advantage of to reduce the computation work and use only the lower left triangle and the main diagonal in the stiffness matrix K.

(1) $K[i,j] := a(\phi_j, \phi_i)_\Omega \quad i = 1, 2, .., M;\ j = 1, 2, .. i$ (a)

(2) $G[i] := G(\phi_i)_\Omega \quad i = 1, 2, .. M$ (b)

(3) $K[i,j] := K[i,j] + a(\phi_j, \phi_i)_{\Gamma_1},\ i = 1, .. M;\ j = 1, .. i$ (c)

(4) $G[i] := G[i] + G(\phi_i)_{\Gamma_1} \quad i = 1, 2, .., M$ (d)

(5) $\forall j \in J_0$

$$\left.\begin{array}{l} G[i] := G[i] - \gamma(N_j) K[j,i] \\ K[j,i] := 0 \end{array}\right\} \forall i \in J,\ i < j$$

$$\left.\begin{array}{l} G[i] := G[i] - \gamma(N_j) K[i,j] \\ K[i,j] := 0 \end{array}\right\} \forall i \in J,\ i > j \quad \text{(e)}$$

$$K[j,i] := 0 \quad \forall i \in J_0,\ i < j$$

$$G[j] := \gamma(N_j); \quad K[j,j] := 1 \quad (5.2)$$

When these relations are written for a storage scheme designed for a sparse matrix which is symmetric, there is reduction in storage.

$$K_{ij} = a(\phi_j, \phi_i)_\Omega = a(\phi_i, \phi_j)_\Omega = \iint_\Omega F(\phi_i, \phi_j)\, dx\, dy \qquad (53)$$

$$i, j = 1, 2, \dots M$$

$$F(\phi_i, \phi_j) = p(\phi_i)_x (\phi_j)_x + p(\phi_i)_y (\phi_j)_y + q\phi_i\phi_j \qquad (54)$$

$K_{ij} \neq 0$ only if nodes N_i and N_j belong to a common element. The primitive matrix K of order $M \times M$ is sparse if the mesh is fine. Decompose K showing the contributions from the individual elements.

$$K = \sum_{l=1}^{L}{}' \hat{K}^{(l)} \quad \text{with} \quad \hat{K}^{(l)}_{ij} = \iint_{e_l} F(\phi_i, \phi_j)\, dx\, dy \qquad (i, j = 1, \dots M) \qquad (55)$$

$\hat{K}^{(l)}$ is even sparser than K as there are only T nodes in an element. With each e_l we may associate a $T \times T$ matrix $K^{(l)}$ with its elements

$$K^{(l)}_{rs} = \iint_{e_l} F(\phi^{(l)}_r, \phi^{(l)}_s)\, dx\, dy, \quad r, s = 1, \dots T \qquad (56)$$

$K^{(l)}$ = the element stiffness matrix (of e_l).

We may then "assemble" the matrix K from the entries of the element matrices $K^{(l)}$. Since both these are symmetric, we need assemble only the lower left triangle. This can be done by a suitable pseudo-ALGOL code.

The elements $K^{(l)}_{rs}$ can be evaluated by rendering the transformation of the element ~~[illegible]~~ onto the

standard element $\tilde{e}$ and then using a standard numerical integration formula.

$$I = \iint_{\tilde{e}} \tilde{\theta}(\tilde{x}, \tilde{y})\, d\tilde{x}\, d\tilde{y} \tag{57}$$

and
$$I_A = \sum_{m=1}^{Q} w_m \tilde{\theta}(\tilde{x}^{(m)}, \tilde{y}^{(m)}) \tag{58}$$

where the integration points $(\tilde{x}^{(m)}, \tilde{y}^{(m)})$ $(m = 1, \cdots Q)$ and the corresponding weights are independent of the function $\tilde{\theta}$.

The evaluation of the element vectors G_i, the edge integrals and arrays in which node $j \in J_0$ can all be reduced / handled by standard numerical procedure.

Properties of the Element Matrices

$$\begin{aligned} -(p u_x)_x - (p u_y)_y + q u &= g \quad \text{in } \Omega \\ u &= r \quad \text{on } \Gamma_0 \\ u_\nu + \sigma u &= \xi \quad \text{on } \Gamma_1 \end{aligned} \tag{59}$$

$$\begin{aligned} & p, q \in C(\bar{\Omega}), \quad \sigma \in C(\Gamma_1), \\ & 0 < p_0 \le p(x,y) \le p_1 \quad \forall (x,y) \in \bar{\Omega} \\ & 0 \le q(x,y) \le q_1 \quad \forall (x,y) \in \bar{\Omega} \\ & 0 < \sigma_0 \le \sigma(x,y) \le \sigma_1 \quad \forall (x,y) \in \Gamma_1 \end{aligned} \tag{60}$$

For any $u, v \in H^1(\Omega)$, define

$$F(u,v) = p u_x v_x + p u_y v_y + q u v \tag{61}$$

$$K_{rs}^{(l)} = \iint_{e_l} F(\phi_r^{(l)}, \phi_s^{(l)})\, dx\,dy, \quad r, s = 1, 2, \cdots, T \tag{62}$$

$K_{rs}^{(l)}$ is symmetric ($= K_{sr}^{(l)}$). Let

$$V_T^{(l)} = \text{SPAN}\{\phi_1^{(l)}, \phi_2^{(l)}, \ldots \phi_T^{(l)}\} \tag{63}$$

$$u^{(l)}(x,y) = \sum_{r=1}^{T}{}' \alpha_r^{(l)} \phi_r^{(l)}(x,y)$$

~~If $\bar{\alpha}^l = [1, 1, \cdots 1]^T$~~ = a column of ~~1's with~~

~~T rows,~~ we ~~have~~

$$(\bar{\alpha}^l)^T K^{(l)} \alpha^l = \iint_{e_l} \{p(u_x^{(l)})^2 + p(u_y^{(l)})^2 + q(u^{(l)})^2\}\, dx\,dy \tag{64}$$

$$(K^{(l)}\alpha^{(l)})_r = \iint_{e_l} \{p(\phi_r^{(l)})_x u_x^l + p(\phi_r^{(l)})_y u_y^l + q\,\phi_r^{(l)} u^{(l)}\}\, dx\,dy \tag{65}$$

Choice of $\bar{\alpha}^l = (1, 1, \cdots 1)^T$ leads to

$u^l(x,y) = 1 \quad \forall (x,y) \in e_l$ and

$$\sum_{r,s=1}^{T}{}' K_{rs}^{(l)} = \iint_{e_l} q\, dx\,dy, \quad \sum_{s=1}^{T}{}' K_{rs}^{(l)} = \iint_{e_l} q\,\phi_r^l\, dx\,dy \tag{66}$$

<u>Theorem</u> If $q(x,y) = 0 \ \forall (x,y) \in e_l$, then $K^{(l)}$ is positive semidefinite and $\sum_{s=1}^{T}{}' K_{rs}^{(l)} = 0$, $r = 1, 2, \cdots, T$. If $q(x,y) > 0$ for some $(x,y) \in e_l$, then $K^{(l)}$ is positive definite. If the basis functions are piecewise linear or piecewise bilinear, $\sum_{s=1}^{T} K_{rs}^{(l)} > 0$, $r = 1, 2, \cdots, T$.

(7) ERROR ANALYSIS

Assume that the problem is posed exactly. There are inevitable errors connected with the formulation of a problem as a model of scientific or social processes and these are neglected. Four types of errors need to be looked into.

(a) Approximation error.

Solution of an exact problem by seeking an approximation to the solution. Element $u_* \in V$ is approximated by $u_*^{(n)} = p_n v^{(n)}$ belonging to V and the error $u_* - p_n v^{(n)}$ is measured by the norm

$$\rho_n := \| u_* - p_n v^{(n)} \| = \| u_* - u_*^{(n)} \|$$

(b) Perturbation error

$\{ x \in X : x \, r \, f \}$ For a given f, find x such that (i) x belongs to a given class X of objects and (ii) x is in a relation r to f

To every positive integer n, an object $f^{(n)}$, a class of objects X_n, an operator $p_n : X_n \to X$ and a relation r_n are assigned. The problem is replaced by

$$\{ v^{(n)} \in X_n : v^n \, r_n \, f^{(n)} \}$$

$f^{(n)}$ is computed from the data of the problem and the computations involve an error. For $f^{(n)}$, we actually use a perturbed version $\tilde{f}^{(n)}$. Likewise $\tilde{r}_n$ for r_n.

A sequence of 'perturbed' approximating problems

$$\{ z^{(n)} \in X_n : \; z^n \; \tilde{r}_n \; \tilde{f}^{(n)}_* \}$$

X_n - perturbation error : $\tilde{\xi}_n := \| z^{(n)}_* - v^{(n)}_* \|_{X_n}$

Perturbation error $\xi_n := \| p_n z^{(n)}_* - p_n v^n_* \|_X$

(3) Algorithm Error

There is some algorithm that produces after a finite number of operations an exact value $z^{(n)}_*$ or some element $w^{(n)} \in X_n$ and the error is $\eta_n := \| p_n w^{(n)} - p_n z^{(n)}_* \|$

X_n - algorithm error $\tilde{\eta}_n := \| w^{(n)} - z^{(n)}_* \|_{X_n}$.

(4) Rounding error

Algorithms operations and computations of functions involve rounding errors. Instead of $w^n \in X_n$, we obtain $\tilde{w}^n \in X_n$.

$\tilde{\zeta}_n := \| \tilde{w}^n - w^n \|_{X_n}$ is the X_n - rounding error.

$\zeta_n := \| p_n \tilde{w}^n - p_n w^n \|_X$: rounding error.

$$\delta_n = \| x_* - p_n \tilde{w}^{(n)} \| \leq e_n + \xi_n + \eta_n + \zeta_n .$$

ERROR ANALYSIS

Second order problems, polygonal domains

$$-(p u_x)_x - (p u_y)_y + q u = g \quad \text{in } \Omega \qquad (1)$$

$$u = \gamma \quad \text{on } \Gamma_0 \qquad (2)$$

$$u_\nu + \sigma u = \xi \quad \text{on } \Gamma_1 \qquad (3)$$

$$H^1_\gamma(\Omega) = \{u \in H^1(\Omega) ; u = \gamma \text{ on } \Gamma_0\} \qquad (4)$$

$$H^1_0(\Omega) = \{v \in H^1(\Omega) ; v = 0 \text{ on } \Gamma_0\} \qquad (5)$$

$$\tilde{V}_N = \{u \in H^1_\gamma(\Omega) ; u = v + u^*, v \in V_N\} \qquad (6)$$

u^* is an arbitrary but fixed function in $H^1_\gamma(\Omega)$ and V_N is an N-dimensional subspace of $H^1_0(\Omega)$. $\hat{u}_N \in \tilde{V}_N$ and is the Ritz-Galerkin approximation to the exact solution $\hat{u} \in H^1_\gamma(\Omega)$. We have the error estimate

$$\| \hat{u} - \hat{u}_N' \|_1 \leq c \min_{u \in \tilde{V}_N} \| \hat{u} - u \|_1 \qquad (7)$$

where c is a constant $(= \sqrt{\rho^{-1}\beta}$ and ρ, β are positive constants that arise in the statements of boundedness and coerciveness of the bilinear form $a(\cdot,\cdot)$). (7) gives an estimate for the discretization error.

The finite element basis functions are $\phi_1, \phi_2, \ldots, \phi_M$ and $V_N = \text{SPAN}\{\phi_i ; i \in J\}$

$$u^*_N(x,y) = \sum_{i \in J_0} \gamma(N_i)\, \phi_i(x,y), \quad (x,y) \in \bar{\Omega}. \qquad (8)$$

We presume that $u_N^* \in H_\gamma^1(\Omega)$ though this is true only for very special boundary functions γ. Also that all integrals in the element matrices and vectors are determined exactly. Since $u_N^* \in H_\gamma^1(\Omega)$,

$$u \in \tilde{V}_N \iff u = \sum_{i=1}^{M} \alpha_i \phi_i \qquad (9)$$

and the coefficients α_i $(i \in J)$ are arbitiary while $\alpha_i (i \in J_0) = \gamma(N_i)$.

$\hat{u}_N^* =$ Ritz - Galerkin approximation $= \hat{u}_N$ where the α_i $(i \in J)$ are the solution of the Ritz- Galerkin equations. $\hat{u}_I =$ the V_M interpolant of $\hat{u}$.

$$V_M = \text{SPAN}\{\phi_1, \ldots, \phi_M\}, \quad \hat{u}_I = \sum_{i=1}^{M} \hat{u}(N_i)\,\phi_i \in V_M \qquad (10)$$

We make the weak assumption that $\hat{u} \in C(\bar{\Omega})$ which holds certainly if $\hat{u} \in H_\gamma^1(\Omega) \cap H^2(\Omega)$. For $i \in J_0$

$\hat{u}_I(N_i) = \hat{u}(N_i) = \gamma(N_i)$ and $\hat{u}_I \in \tilde{V}_N$

$$\|\hat{u} - \hat{u}_N\|_1 \leq c\,\|\hat{u} - \hat{u}_I\|_1 \qquad (11)$$

and to obtain a bound on the discretization error $\hat{u} - \hat{u}_N$ we may examine the interpolation error $\hat{u} - \hat{u}_I$.

<u>An example</u> (A simple one-dimensional problem).

$[a, b]$; A finite element mesh with nodes at x_i.

$$a = x_0 = x_1 < x_2 < \cdots < x_M = x_{M+1} = b$$

Elements $[x_i, x_{i+1}]$ $i = 1, 2, \ldots, M-1$

Define the basis functions

$$\phi_i(x) = \begin{cases} (x - x_{i-1})/h_{i-1} & (x_{i-1} \le x < x_i) \\ (x_{i+1} - x)/h_i & (x_i \le x \le x_{i+1}) \\ 0 & \text{otherwise} \end{cases} \quad (12$$

$h_i = x_{i+1} - x_i$. The V_m-interpolant of u is u_1 which is a continuous, piecewise linear function, given by

$$u_1(x) = \sum_{i=1}^{M}{}' u(N_i)\,\phi_i(x) \qquad (13)$$

Let $u \in H^2(a,b)$. The Sobolev seminorm is

$$|u|_2 = \left(\int_a^b (u''(x))^2\, dx \right)^{1/2} \qquad (14)$$

u is necessarily $C^1[a,b]$ and by the mean value theorem, there exists a ξ in (x_i, x_{i+1}) such that

$$u'(\xi) = h_i^{-1}[u(x_{i+1}) - u(x_i)]$$

$u_1'(x) = u'(\xi)$ $\forall$ $x \in (x_i, x_{i+1})$ and in this interval

$u'(x) - u_1'(x) = u'(x) - u'(\xi) = \int_\xi^x u''(t)\,dt$. By Schwarz's inequality

$$|u'(x) - u_1'(x)| \le \left(\int_\xi^x (1)^2\, dt \right)^{1/2} \left(\int_\xi^x (u''(t))^2 dt \right)^{1/2}$$

$$\le (h_i^{1/2}) \int_{x_i}^{x_{i+1}} (u''(t))^2 dt \quad , x \in (x_i, x_{i+1}) \quad (15)$$

$$u(x) - u_1(x) = \int_{x_i}^{x} [u'(t) - u_1'(t)]\,dt$$

$$|u(x) - u_1(x)| \leq \int_{x_i}^{x_{i+1}} |u'(t) - u_1'(t)|\,dt \leq h_i^{3/2}\left(\int_{x_i}^{x_{i+1}} (u''(t))^2\,dt\right)^{1/2}$$

$$x \in [x_i, x_{i+1}] \qquad (16)$$

Squaring (15), integrating over the element and then summing over all the elements, we get

$$|u - u_1|_1 \leq h\,|u|_2 \qquad (17)$$

Similarly from (16) we get $|u - u_1|_0 \leq h^2\,|u|_2$ (18)

with h in (17), (18) $=$ max h_i $(i = 1, 2, \ldots, m-1)$.

$$\|u - u_1\|_1^2 = |u - u_1|_0^2 + |u - u_1|_1^2$$

and hence $\|u - u_1\|_1 \leq c\,h\,|u|_2$ (19)

with $c = \{1 + (b-a)^2\}^{1/2}$

With more refinement of analysis we can derive the result that for any $u \in H^2(a,b)$,

$$|u - u_1|_1 \leq \frac{h}{\pi}|u|_2, \quad |u - u_1|_0 \leq \frac{h^2}{\pi^2}|u|_2$$

and these are the strongest possible bounds.

The indices k and m were defined earlier to be the largest integers such that

$$P_k(\bar{\Omega}) \subseteq V_m \subset C^{m-1}(\bar{\Omega}) \qquad (20)$$

and this implies that $V_m \subset H^m(\Omega)$. In the example $k = m = 1$. The seminorm bounds are of the

form $\quad |u - u_1|_s \leq c_s h^{K+1-s} |u|_{K+1} \quad (s = 0, 1, \ldots, m)$

where $c_0, c_1, \ldots, c_m$ are independent of u.

Interpolation error: Two dimensional finite elements

e_l = a typical element. For any $\theta \in H^s(e_l)$

$|\theta|_{s, e_l}$ is the Sobolev seminorm of order s.

$$|u - u_1|_s^2 = \sum_{l=1}^{L} |u - u_1|_{s, e_l}^2 \tag{21}$$

Essential theorems

(1) Let there be given a standard element $\tilde{e}$ and let standard local basis functions $\tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_T$ and let K be the largest integer such that

$$P_K(\tilde{e}) \subseteq \text{SPAN}\{\tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_T\} \tag{22}$$

$P_K(\tilde{e})$ is the space of all polynomials of degree $\leq K$ defined on $\tilde{e}$. There exists a constant $\hat{c}$ independent of $\hat{u}$ such that

$$|\tilde{u} - \tilde{u}_1|_{s, \tilde{e}} \leq \hat{c}\, |\tilde{u}|_{K+1, \tilde{e}} \quad \forall\ \tilde{u} \in H^{K+1}(\tilde{e}) \tag{23}$$

$$s = 0, 1, 2, \ldots, K+1.$$

(2) Let $\bar{\Omega} \subset R^2$ be a given polygonal domain and let there be a finite element mesh on $\bar{\Omega}$ with triangular elements, and let the associated basis functions be $\phi_1, \phi_2, \ldots, \phi_M$.

h = greatest edge length, α = smallest angle in the mesh.

K = largest integer for which $P_K(\bar{\Omega}) \subseteq V_M$. Then there exists a constant c, independent of u and

the mesh, such that

$$|u - u_I|_s \leq c (\sin \alpha)^{-s} h^{K+1-s} |u|_{K+1} \qquad (24)$$

$$\forall\ u \in H^{K+1}(\Omega), \quad s = 0, 1$$

where u_I is the V_M interpolant of u.

The V_M interpolant is of the Lagrange type, viz.,

$u_I(x,y) = \sum_{i=1}^{M} u(N_i)\, \phi_i(x,y)$. The first order derivatives of the global basis functions and hence of u_I have step discontinuities across element boundaries and in general $u_I \notin H^2(\Omega)$. The Sobolev seminorm $|u - u_I|_s$ must be restricted to the values $s = 0, 1$.

Rectangular Elements.

Let $\bar{\Omega}$ be a given axiparallel domain $\subset R^2$ and let there be a finite element mesh on $\bar{\Omega}$ with axiparallel rectangular elements and let the associated basis functions be $\phi_1, \phi_2, \ldots, \phi_M$. Let h and β be the largest edge length and smallest edge ratio (= smallest edge h^* / largest edge h). K = largest integer for which $P_K(\bar{\Omega}) \subseteq V_M$. Then there exists a value c, independent of u and the mesh, such that

$$|u - u_I|_s \leq c\, \beta^{-s} h^{K+1-s} |u|_{K+1} \quad \forall\ u \in H^{K+1}(\Omega), \qquad (25)$$

$s = 0, 1$

where u_I is the V_M interpolant of u.

A cardinal rule regarding the generation of a finite element mesh: avoid small angles in a triangular

mesh and small edge ratios in a rectangular mesh.

Let $\{M_p\}$ $(p = 1, 2, \ldots\ldots)$ be an infinite sequence of meshes on $\bar{\Omega}$ which is a regular family of either triangular or rectangular elements on $\bar{\Omega}$, such that $h \to 0$ as $p \to \infty$. The family is regular if all the angles in the triangles are bounded below by some $\alpha > 0$ or in the case of rectangular elements, if all the edge ratios are bounded below by some $\beta > 0$. If $u \in H^{K+1}(\Omega)$ and u_I is the corresponding interpolant derived from all the meshes in a regular family,

$$|u - u_I|_s \leq \tilde{c}\, h^{K+1-s} |u|_{K+1}, \quad h \to 0 \text{ for } s = 0, 1. \qquad (26)$$

If $u \in H^{p+1}(\Omega)$ for some $p < K$, there is a constant c^* such that

$$|u - u_I|_s \leq c^* h^{p+1-s} |u|_{p+1}, \quad s = 0, 1. \qquad (27)$$

Discretization error

The discretization error $\|\hat{u} - \hat{u}_N\|_1$ and the interpolation error $\|\hat{u} - \hat{u}_I\|_1$ are related through the inequality (28)

$$\|\hat{u} - \hat{u}_N\|_1 \leq c\,\|\hat{u} - \hat{u}_I\|_1$$

using the identity $\|\hat{u} - \hat{u}_I\|_1^2 = |\hat{u} - \hat{u}_I|_0^2 + |\hat{u} - \hat{u}_I|_1^2$ and the bounds in (26) above with $s = 0$, $s = 1$ we see

that $\|\hat{u} - \hat{u}_N\|_1 \leq \tilde{c}_1 h^K |\hat{u}|_{K+1}$ (29)

for some constant $\tilde{c}_1$. If $\hat{u} \in H^{p+1}(\Omega)$ and $p < K$, from (28) we see that $\|\hat{u} - \hat{u}_N\|_1 \leq c_1^* h^p |\hat{u}|_{p+1}$ for some constant c_1^*. From (29) we have the seminorm bounds

$$|\hat{u} - \hat{u}_N|_0 \leq \tilde{c}_1 h^K |\hat{u}|_{K+1}, \quad |\hat{u} - \hat{u}_N|_1 \leq \tilde{c}_1 h^K |\hat{u}|_{K+1} \quad (30)$$

The gradient error $|\hat{u} - \hat{u}_N|_1$ has a bound (order h^K) agreeing with the bound (26) with $s = 1$. The bound on $|\hat{u} - \hat{u}_1|_0$ in (26) is $O(h^{K+1})$ while that in (30) is $O(h^K)$, a weak result in comparison.

Pointwise Error Estimates

Discretization error $\hat{u}(x) - \hat{u}_N(x)$, $x \in \bar{\Omega}$.

$$\text{Maximum norm of the error} = \|\hat{u} - \hat{u}_N\|_\infty = \max_{x \in \bar{\Omega}} |\hat{u}(x) - \hat{u}_N(x)| \quad (31)$$

We may employ the Green's function of a boundary value problem to analyze the behaviour of the discretization error.

Illustration: A two-point boundary value problem,

$$\left.\begin{aligned} -(p(x)u')' + q(x)u &= g(x) \quad -1 < x < 1 \\ u(-1) &= u(1) = 0. \end{aligned}\right\} \quad (32)$$

$p \in C^1[-1, +1]$, $q, g \in C[-1, 1]$, $0 < p_0 \leq p(x) \leq p_1$, $0 \leq q(x) \leq q_1$.

Let $K(x,\xi)$ be the Green's function of the problem

and $a(u,v) = \int_{-1}^{+1} (p\,u'v' + q\,uv)\,d\xi, \quad u,v \in \mathring{H}^1(-1,+1)$

Then $u(x) = a(K(x,\cdot), u(\cdot)) \quad \forall\, u \in \mathring{H}^1(-1,1)$.

The Ritz-Galerkin approximation $\hat{u}_N \in V_N$ satisfies

$$a(u, \hat{u} - \hat{u}_N) = 0 \quad \forall\, u \in V_N. \tag{33}$$

$$\hat{u}(x) - \hat{u}_N(x) = a(K(x,\cdot), \hat{u}(\cdot) - \hat{u}_N(\cdot))$$

$$= a([K(x,\cdot) - u], \hat{u}(\cdot) - \hat{u}_N(\cdot)) \quad \forall\, u \in V_N$$

The bilinear form is bounded.

$$|a(u,v)| \le \beta \|u\|_1 \|v\|_1 \quad \forall\, u,v \in \mathring{H}^1(-1,1).$$

Also from (26) above (which is valid also for this one-dimensional problem)

$$\|\hat{u} - \hat{u}_N\|_1 = O(h^K) \quad \text{if} \quad \hat{u} \in H^{K+1}(-1,1)$$

Let the V_M interpolant of $K(x,\xi)$, $-1 \le \xi \le 1$ be $K_1(x,\xi)$. $K(x,1) = K(x,-1) = 0$ by the definition of the Green's function. We see that $K_1(x,\xi) \in V_N$.

$$|\hat{u}(x) - \hat{u}_N(x)| = |a(K(x,\cdot) - K_1(x,\cdot), \hat{u}(\cdot) - \hat{u}_N(\cdot))|$$

$$= O(h^K)\, \|K(x,\cdot) - K_1(x,\cdot)\|_1 \tag{34}$$

The Green's function belongs to $\mathring{H}(-1,1)$ and is smooth in the two subintervals $-1 \le \xi < x$ and $x \le \xi \le 1$ but there is a jump discontinuity at $\xi = x$. Thus $u \notin H^2(-1,1)$. From (27)

$$\|K(x,\cdot) - K_1(x,\cdot)\|_1 = O(h^0) \tag{35}$$

From (34), (35), we get $|\hat{u}(x) - \hat{u}_N(x)|_\infty = O(h^K)$. (36)

We know that $\| \hat{u} - \hat{u}_N \| = O(h^{K+1})$ and the above result is not of significance. A more refined analysis leads to the conclusion $\| \hat{u} - \hat{u}_N \|_\infty = O(h^{K+1})$

<u>Pointwise error at a node</u> $x = x_i$.

The jump discontinuity in the first derivative of the interpolation error $K(x_i, \xi)$ has no effect on the interpolation error because it occurs at a node. $K(x_i, \xi)$ in $-1 < \xi < x_i$ and $x_i < \xi < 1$ belong to $H^{K+1}(-1, x_i)$ and $H^{K+1}(x_i, 1)$ respectively and decomposing the interpolation error into contributions from the two subintervals, we have from (26),

$$\| K(x_i, \cdot) - K_1(x_i, \cdot) \|_1 = O(h^K) \quad (36)$$

Hence $\quad | \hat{u}(x_i) - \hat{u}_N(x_i) | = O(h^{2K}) \quad (37)$

on using (34) above. If $K > 1$, this is asymptotically faster than $O(h^{K+1})$ which is the order of convergence of $\| \hat{u} - \hat{u}_N \|_\infty$. This is referred to as <u>superconvergence</u>.

Maximum-norm and pointwise error analysis for second order boundary value problems in 2 or more dimensions is complicated since the Green's function does not belong to $H^1(\Omega)$.

For the boundary value problem $-\nabla^2 u + u = g$ in $\Omega \subset R^2$, $u_\nu = 0$ on Γ, using triangular elements, it is established that

$$\| \hat{u} - \hat{u}_N \|_\infty \leq \begin{cases} ch^2 |\log h| & \text{for } K = 1 \\ ch^{K+1} & \text{for } K \geq 2 \end{cases}$$

CONDITION NUMBERS

H is an $N \times N$ positive definite matrix with the eigen values $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$. Then $\kappa(H) = \lambda_N / \lambda_1$ is the spectral condition number of H. In the Ritz-Galerkin equations

$$K\alpha = G \tag{38}$$

the stiffness matrix K is positive definite, when the bilinear form $a(\cdot,\cdot)$ is symmetric and coercive. If the number $\kappa(K)$ is large, the system (38) is ill-conditioned and even a small change in G or K may cause a large change in the solution α and then the computed solution of (38) is unreliable. A large condition number also reduces the rate of convergence of the iterative methods of solving the system (38). It is important to assess the magnitude of $\kappa(K)$.

Consider a boundary value problem of second order in n space ~~variables~~ dimensions and let the bilinear form $a(\cdot,\cdot)$ be symmetric, bounded and coercive.

$$\rho \|u\|_1^2 \leq a(u,u) \leq \beta \|u\|_1^2 \quad \forall u \in V \tag{39}$$

$\mathring{H}^1(\Omega) \subseteq V \subseteq H^1(\Omega)$. Solution $\hat{u}$ belongs to V unless there is an inhomogeneous boundary condition of Dirichlet's type. In the latter case $\hat{u} = \hat{v} + u^*$ for $v \in V$ and u^* satisfies Dirichlet's condition. Without loss of generality we take the N nodes where $\hat{u}$ is unknown

in the first N orders. $K = K_{ij}$

$$K_{ij} = a(\phi_j, \phi_i) \quad (i, j = 1, 2, \cdots N.) \tag{40}$$

Mass matrix $m_{ij} = (\phi_i, \phi_j) = \int_\Omega \phi_i \phi_j \, d\Omega$ (41)

$(i, j = 1, \cdots N.)$

$M^{(l)}$ $(l = 1, \cdots L)$ are element mass matrices and

$$m_{rs}^{(l)} = \int_{e_l} \phi_r^{(l)}(x) \, \phi_s^{(l)}(x) \, d\Omega \quad r, s = 1, \cdots T \tag{42}$$

$$\alpha^T M \alpha = \int (u(x))^2 \, d\Omega, \quad u(x) = \sum_{i=1}^{N} \alpha_i \phi_i(x)$$

M and similarly ~~every~~ all $M^{(l)}$'s are positive definite.

$$\left.\begin{array}{ll} \text{Eigenvalues of } M: & 0 < \eta_1 \le \eta_2 \le \cdots \le \eta_N \\ \text{Eigenvalues of } M^{(l)}: & 0 < \eta_1^{(l)} \le \eta_2^{(l)} \le \cdots \le \eta_T^{(l)} \end{array}\right\} \tag{43}$$

$$\eta_1 = \min_{\alpha \in R^N} \frac{\alpha^T M \alpha}{\alpha^T \alpha}, \quad \eta_N = \max_{\alpha \in R^N} \frac{\alpha^T M \alpha}{\alpha^T \alpha} \tag{44}$$

$$\eta_1^{(l)} = \min_{\beta \in R^T} \frac{\beta^T M^l \beta}{\beta^T \beta} \quad \eta_T^{(l)} = \max_{\beta \in R^T} \frac{\beta^T M^l \beta}{\beta^T \beta} \tag{45}$$

$$\alpha^T M \alpha = \sum_{l=1}^{L}{}' (\alpha^l)^T M^l \alpha^l \tag{46}$$

$$\alpha_r^{(l)} = \begin{cases} \alpha_{i_r^{(l)}} & \text{if } 1 \le i_r^{(l)} \le N \\ 0 & \text{if } i_r^{(l)} > N \end{cases} \quad r = 1, \cdots T \tag{47}$$

$$\alpha^T M \alpha \geq \min_{1 \le l \le L} \eta_1^l \sum_{l=1}^{L}{}' \alpha^{(l)T} \alpha^{(l)} \geq p_1 \left(\min_l \eta_1^{(l)}\right) \|\alpha\|^2$$

$$\& \quad \le p_2 \left(\max_l \eta_T^{(l)}\right) \|\alpha\|^2.$$

(13)

p_1 = smallest number of elements to which any of the first N nodes belongs and p_2 = largest number of elements to which any of the first N nodes belongs.

$$p_1 \min_{(l)} \eta_1^{(l)} \le \eta_1 \le \eta_N \le p_2 \max_{(l)} \eta_T^{(l)} \qquad (48)$$

The mapping of the standard element $\tilde{e}$ onto e_l is affine and $\tilde{D}$ is the Jacobian of the mapping and $\tilde{\phi}_r$ ($r = 1, 2, \cdots T$) are the standard local basis functions. A_l and $\tilde{A}$ are the measures (areas / lengths / volumes depending on n) and

$$|\tilde{D}| = A_l / \tilde{A} .$$

$$M^{(l)} = (A_l / \tilde{A}) \tilde{M}$$

$$p_1 (\tilde{\eta}_1 / \tilde{A}) \min A_l \le \eta_1 \le \eta_N \le p_2 (\tilde{\eta}_T / \tilde{A}) \max A_l \qquad (49)$$

Apply the above key result to an infinite family of finite element meshes on $\bar{\Omega}$ having the property that there exist constants c_0, c_1, c_2 and c_3 such that for every mesh

$$\left. \begin{aligned} p_2 \le c_0 , \quad & c_1 N^{-1} \le A_l \le c_2 N^{-1}, \ (l = 1, \cdots L) \\ c_3 N^{-\frac{1}{n}} & \le h_i \le c_4 N^{-1/n} \end{aligned} \right\} \qquad (50)$$

Then constants c_5, c_6 exist such that

$$c_5 N^{-1} \le \eta_1 \le \eta_N \le c_6 N^{-1} \qquad (51)$$

$$a(u, u) = \alpha^T K \alpha , \quad \|u\|^2 = \alpha^T M \alpha \quad \text{with } u(x) = \sum_{i=1}^{N} \alpha_i \phi_i(x)$$

$$\lambda_1 = \min_{\alpha \in R^N} \frac{\alpha^T K \alpha}{\alpha^T \alpha} \quad \text{and} \quad \frac{\alpha^T K \alpha}{\alpha^T \alpha} \geq \rho \frac{\alpha^T M \alpha}{\alpha^T \alpha}$$

$\forall \alpha \in R^N$. $\quad \lambda_1 \geq \rho \eta_1 \geq \rho\, \phi_1 (\tilde{\eta}_1 / \tilde{A}) \min_{(L)} A_L$

Then $\quad \lambda_1 \geq c_7 N^{-1}, \quad c_7 = \rho c_5$

$$\alpha^T Q \alpha = |u|_1^2, \quad u(x) = \sum_{i=1}^{N} \alpha_i \phi_i(x)$$

Q is positive definite / semi definite. Eigenvalues of Q are

$$0 \leq \xi_1 \leq \xi_2 \leq \cdots \leq \xi_N$$

$$a(u,u) \leq \beta\left(\|u\|^2 + |u|_1^2\right)$$

$$\therefore \quad \lambda_N \leq \beta(\eta_N + \xi_N) \qquad (52)$$

$$\eta_N \leq c_6 N^{-1}, \quad \xi_N \leq c_8 N^{\frac{2}{n}-1}$$

$$\lambda_N \leq c_9 N^{\frac{2}{n}-1}$$

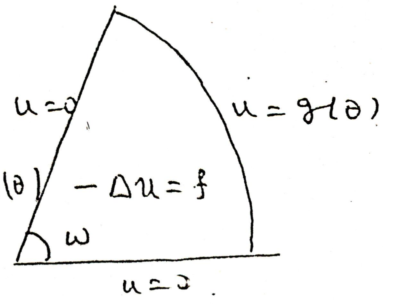
$$\kappa(K) \leq c_{10} N^{\frac{2}{n}}, \quad c_{10} = c_9/c_7 \qquad (53)$$

The bound on $\kappa(K)$ shows that for a fixed large value of N, the stiffness matrix of a three dimensional problem $(n = 3)$ is better conditioned than that of a two dimensional problem and this in turn is better conditioned than that of a one-dimensional problem. The bound is independent of the accuracy of the finite element basis functions, i.e. is independent of k = the greatest integer for which $P_k(\bar{\Omega}) \subset V_M \subset C^{m-1}(\bar{\Omega})$, $P_k(\bar{\Omega})$ being the space of all polynomials of degree $\leq k$.

For 4th order problems $\kappa(K) = O(N^{4/n})$.

Singularities

$$\Delta u = u_{rr} + \frac{1}{r} u_r + \frac{1}{r^2} u_{\theta\theta}$$



$$\hat{u}(r,\theta) = S_1(r,\theta) + S_2(r,\theta)$$

$$= \sum_{j=1}^{\infty} g_j \, r^{\nu_j} \psi_j(\theta) + \sum_{j=1}^{\infty} \sum_{l=0}^{\infty} f_{jl} b_{jl} \, x_l(r) \psi_j(\theta)$$

$$\nu_j = \frac{\pi j}{\omega}, \quad \psi_j(\theta) = \sqrt{\frac{2}{\omega}} \sin \nu_j \theta$$

S_1 and S_2 are continuous in $\bar{\Omega}$ but partial derivatives with respect to r involve singularities at the origin. The terms which introduce the singularities for the first time are most important and may be termed the 'critical terms'.

Critical Terms in S_1, S_2 and $\hat{u}$

case	$\alpha = \frac{\pi}{\omega}$	critical terms in S_1	critical terms in S_2	critical terms in $\hat{u}$
1	irrational	$r^\alpha \sin \alpha\theta$	none	$r^\alpha \sin \alpha\theta$
2	$n/m \quad (m>1)$	$r^\alpha \sin \alpha\theta$	$r^n (\log r) \sin n\theta$	$r^\alpha \sin \alpha\theta$
3	$n \quad (n>1)$	none	$r^\alpha (\log r) \sin \alpha\theta$	$r^\alpha \log r \sin \alpha\theta$

In case (2) $\alpha = \frac{n}{m} < n$ and then $\lim_{r \to 0} \left(r^\alpha / r^n \log r \right) = \infty$

Thus the term $r^\alpha \sin \alpha\theta$ dominates.

In all the cases the first singular derivative of the critical term of $\hat{u}$ is of order s which is the smallest integer satisfying the condition $> \alpha$. Let $\alpha = \frac{\pi}{\omega}$

We want to find the largest integer p that makes $\hat{u} \in H^p(\Omega)$ or equivalently $|\hat{u}|_p < \infty$

$F(x,y) = \sum\limits_{|\nu|=p}' (\hat{u}_c^\nu)^2$ (critical terms only are retained). $\iint\limits_\Omega F(x,y)\,dx\,dy = \iint\limits_\Omega \tilde{F}(r,\theta)\, r\,dr\,d\theta$

$\hat{u}_c$ denotes the critical terms in $\hat{u}$ and ν is the multi-index notation for partial differentiation.

$|\hat{u}_c|_p < \infty \iff \int_0^R r^{2\alpha-2p+1}\,dr < \infty$ (cases 1, 2)

$\int_0^R r^{2\alpha-2p+1} (\log r)^2\,dr < \infty$ (case 3)

$s =$ the largest such integer p.

$\pi \leq \omega \leq 2\pi \quad (\alpha \leq 1) \implies \hat{u} \in H^1(\Omega)$

$\frac{\pi}{2} \leq \omega < \pi \quad (\alpha \leq \frac{1}{2}) \implies \hat{u} \in H^2(\Omega)$

$\frac{\pi}{3} \leq \omega < \frac{\pi}{2} \quad (\alpha \leq \frac{1}{3}) \implies \hat{u} \in H^3(\Omega)$

Thus the order of the Sobolev space seems to be a step function with respect to α or ω, whereas intuitively one expects "smoothness" ~~to~~ be a continuous feature of ω. To circumvent this situation one may introduce the concept of a fractional Sobolev space $H^t(\omega)$, t being any real number. For any $\alpha > 0$,

$\omega = \frac{\pi}{\alpha} \implies \hat{u} \in H^{\alpha+1-\epsilon}(\Omega)$, $\epsilon > 0$ and small but arbitrary. The solution becomes less smooth as the angle of the corner increases. If $\frac{\pi}{p+1} \leq \omega < \frac{\pi}{p}$ and $p =$ a positive integer, $\|\hat{u} - \hat{u}_N\| \leq c\,h^q |\hat{u}|_{q+1}$ and $q = \min(p, K)$, $K =$ the parameterised basis function

8. One dimensional conductive/convective heat transfer.

Insulated rod of cross section A, length L, conductivity K, convection coefficient β and ambient temperature T_∞, distributed heat source = f, q = heat flux.

$$-\frac{d}{dx}\left(KA\frac{dT}{dx}\right) = f, \quad 0 < x < L \qquad (1)$$

$$T(0) = T_0, \quad \left\{KA\frac{dT}{dx} + \beta(T - T_\infty) + q\right\}_{x=L} = 0 \qquad (2)$$

$$0 = \int_{x_A}^{x_B}\left\{-\frac{d}{dx}\left(KA\frac{dT}{dx}\right) - f\right\} v\,dx = \int_{x_A}^{x_B}\left(KA\frac{dT}{dx}\frac{dv}{dx} - fv\right)dx - \left[\left(KA\frac{dT}{dx}\right)v\right] \qquad (3)$$

$$\text{Let } KA = a = \begin{cases} a_1 & 0 \le x \le L_1 \\ a_2 & L_1 \le x \le L \end{cases} \qquad (4)$$

$$\int_{x_A}^{x_B}\left(a\frac{dT}{dx}\frac{dv}{dx} - fv\right)dx + \left[v\left(-a\frac{dT}{dx}\right)\right]_{x_A}^{x_B} = 0 \qquad (5)$$

$$\int_{x_A}^{x_B}\left(a\frac{dT}{dx}\frac{dv}{dx} - fv\right)dx + \left[v\{\beta(T - T_\infty) + q\}\right]_{x_A}^{x_B} = 0 \qquad (6)$$

{Invoking the natural b.c. $\left[a\frac{dT}{dx} = -\beta(T - T_\infty) - q\right)$ at the limits x_A, x_B}.

(6) motivates the selection of the bilinear form

$$B(T, v) = \int_{x_A}^{x_B} a\frac{dT}{dx}\frac{dv}{dx}\,dx \qquad (7)$$

and (6) may be rewritten in the form

$$B(T, v) - l\,v = 0 \qquad (8)$$

$$l v = \int_{x_A}^{x_B} f v \, dx - \beta v(x_B) T(x_B) + \beta v(x_A) T(x_A)$$

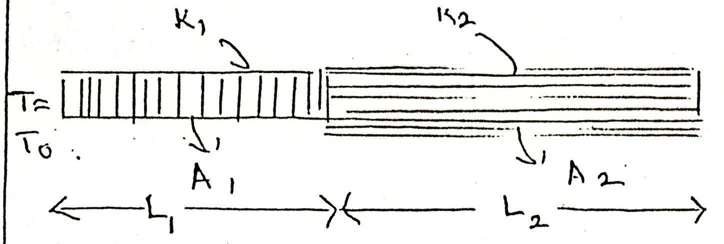
$$+ \beta T_\infty \left(v(x_B) - v(x_A) \right) - q(x_B) v(x_B) + q(x_A) v(x_A).$$

The natural b.c. (the second statement in (2) (9) above) is included in arriving at the variational version of the problem. The associated functional for the element is

$$I_e(T) = \int_{x_A}^{x_B} \left\{ \frac{a}{2} \left(\frac{dT}{dx} \right)^2 - Tf \right\} dx - T(x_A) P_A - T(x_B) P_B \quad \ldots (10)$$

$$\text{with } P_A = \left(- a \frac{dT}{dx} \right)_{x_A}, \quad P_B = \left(+ a \frac{dT}{dx} \right)_{x_B} \quad (11)$$

In (2) T is prescribed at $x = 0$ — an essential b.c. and hence T is the (primary) variable — or degree of freedom — at the nodes of the elements. Statement (6) or (8) along with (9) contains only the first derivative of T and hence linear interpolation is suitable for the problem. Statement (4) indicates a single step change in the parameter $a = KA$ which may arise with change in the cross section area A or the material property — thermal conductivity K or both. This compels the choice of at least two elements. Take two elements.



$L_1 + L_2 = L$. Take $f = 0$.

$$T = c_1 + c_2 x \qquad T_1^{(e)} = T(0) = c_1 \tag{12}$$

$$T_2^{(e)} = T(h_e) = c_1 + c_2 h_e \Rightarrow c_1 = T_1^{(e)}, \; c_2 = \frac{T_2^{(e)} - T_1^{(e)}}{h_e}$$

$$T = T_1^{(e)} \psi_1^{(e)}(x) + T_2^{(e)} \psi_2^{(e)}(x) = \sum_{j=1}^{2} T_j^{(e)} \psi_j^{(e)}(x) \tag{13}$$

$$\psi_1^{(e)}(x) = 1 - \frac{x}{h_e}, \quad \psi_2^{(e)} = \frac{x}{h_e}, \quad 0 \le x \le h_e \tag{14}$$

The variational equation(s) [8], (9) $\Rightarrow$

$$\int_{x_A}^{x_B} a \left(\sum_{j=1}^{2} T_j^{(e)} \frac{d}{dx} \psi_j^{(e)}(x) \right) \frac{dv}{dx} \, dx + \beta v(x_B) T(x_B)$$

$$- \beta v(x_A) T(x_A) = \beta T_\infty \left[v(x_B) - v(x_A) \right]$$

$$- q(x_B) v(x_B) + q(x_A) v(x_A) \tag{15}$$

Taking $v = \psi_i$ in (15) and substituting for $T(x_B)$ and $T(x_A)$ from (13), we have

$$\sum_{j=1}^{2} \left\{ \int_{x_e}^{x_{e+1}} a \frac{d\psi_i}{dx} \frac{d\psi_j}{dx} \, dx + \beta \left[\psi_i(x_{e+1}) \psi_j(x_{e+1}) - \psi_i(x_e) \psi_j(x_e) \right] \right\} T_j$$

$$= \beta T_\infty \left[\psi_i(x_{e+1}) - \psi_i(x_e) \right] + q(x_e) \psi_i(x_e) - q(x_{e+1}) \psi_i(x_{e+1}) \tag{16}$$

$$(K^{(e)})\,(T^{(e)}) = F^{(e)} \qquad (17)$$

$$K_{ij}^{(e)} = \int_{x_e}^{x_{e+1}} a \frac{d\psi_i}{dx}\frac{d\psi_j}{dx}\,dx + \beta\left[\psi_i(x_{e+1})\psi_j(x_{e+1}) - \psi_i(x_e)\psi_j(x_e)\right]$$

$$F_j^{(e)} = \beta T_\infty\left[\psi_j(x_{e+1}) - \psi_j(x_e)\right] + \psi_j(x_e)q(x_e) - \psi_j(x_{e+1})q(x_{e+1})$$

(18),(19)

$$K^{(e)} = \frac{a_e}{h_e}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \begin{pmatrix} -\beta_1^{(e)} & 0 \\ 0 & \beta_2^{e} \end{pmatrix}$$

$$F^{(e)} = T_\infty \begin{pmatrix} -\beta_1^{(e)} \\ \beta_2^{(e)} \end{pmatrix} + \begin{pmatrix} -q_1^{(e)} \\ q_2^{(e)} \end{pmatrix} \qquad (20)$$

$\beta_i^{(e)}$, $q_i^{(e)}$ are the film coefficient and heat flux values at node i of element e.

The two elements are to be assembled.

$$\begin{pmatrix} K^{(1)} & \\ & K^{(2)} \end{pmatrix}\begin{pmatrix} T^{(1)} \\ T^{(2)} \end{pmatrix} = \begin{pmatrix} F^{(1)} \\ F^{(2)} \end{pmatrix} \qquad (21)$$

Element involves two nodes numbered (1), (2) in the global ordering and element two with two nodes that are numbered 2 and 3 in the global ordering. The global node 2 is the transition node from

element 1 to 2.

Take the numerical values for the parameters as follows.

$$(a_1, \quad a_2) = (76, \quad 96) \text{ W-cm}/{}^\circ C$$

$$(L_1, \quad L_2 = L - L_1, \quad L) = (4, \quad 10-4=6, \quad 10) \text{ cm}.$$

$$(T_0, \quad T_\infty) = (100, \quad 30)\,{}^\circ C$$

$$(\beta_1^{(1)}, \quad \beta_2^{(1)}, \quad \beta_1^{(2)}, \quad \beta_2^{(2)}) = (0, \quad 10, \quad 10, \quad 10) \text{ W/cm}^2. \tag{22}$$

At a transition or interelement node there is no convective contribution and $T_\infty \beta_1^{(e+1)} - T_\infty \beta_2^{(e)} = 0$.

$$[K^{(1)}] = 19\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + 10\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 19 & -19 \\ -19 & 29 \end{pmatrix} \tag{23}$$

$$[K^{(2)}] = 16\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + 10\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 6 & -16 \\ -16 & 26 \end{pmatrix} \tag{24}$$

$$T(x) = \begin{cases} \alpha_1 \psi_1^{(1)}(x) + \alpha_2 \psi_2^{(1)}(x) & 0 \le x \le L_1 = 4 \\ \alpha_2 \psi_1^{(2)}(x) + \alpha_3 \psi_2^{(2)}(x) & 4 \le x \le 10 \end{cases} \tag{25}$$

$$F^{(1)} = \begin{pmatrix} -\beta_1^{(1)} T_\infty - q_1^{(1)} \\ \beta_2^{(1)} T_\infty + q_2^{(1)} \end{pmatrix}, \quad F^{(2)} = \begin{pmatrix} -\beta_1^{(2)} T_\infty - q_1^{(2)} \\ 300 + q_2^{(2)} \end{pmatrix} \tag{26}$$

$$\begin{pmatrix} 19 & -19 & 0 \\ -19 & 35 & -16 \\ 0 & -16 & 26 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} -q_1^{(1)} - \beta_1^{(1)} T_\infty \\ (\beta_2^{(1)} - \beta_1^{(2)}) T_\infty + q_2^{(1)} - q_1^{(2)} \\ 300 + q_2^{(2)} \end{pmatrix} \tag{27}$$

Boundary conditions and equilibrium of internal heat flux $\Rightarrow$

$$\alpha_1 = 100^0 C, \quad q_2^{(1)} - q_1^{(2)} = 0, \quad q_2^{(2)} = \hat{q} = 0$$

Hence the system:

$$\begin{pmatrix} 19 & -19 & 0 \\ -19 & 35 & -16 \\ 0 & -16 & 26 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} -q_1^{(1)} \\ 0 \\ 300 \end{pmatrix} \quad (2$$

From the last two rows of this linear system, viz.

$$\begin{pmatrix} 35 & -16 \\ -16 & 26 \end{pmatrix} \begin{pmatrix} \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} 19\alpha_1 \\ 300 \end{pmatrix} = \begin{pmatrix} 1900 \\ 300 \end{pmatrix}$$

we have $\alpha_2 = \dfrac{54200}{654}$, $\alpha_3 = \dfrac{45900}{654}$ (2

From the first row of (28), we have

$$19\alpha_1 - 19\alpha_2 = -q_1^{(1)} \Rightarrow q_1^{(1)} = 19(\alpha_2 - \alpha_1)$$

$$T(x) = \begin{cases} 100\left(1 - \dfrac{x}{4}\right) + \dfrac{54200}{654}\left(\dfrac{x}{4}\right) & 0 \le x \le 4 \\ \dfrac{54200}{654}\left(\dfrac{10 - x}{6}\right) + \dfrac{45900}{654}\left(\dfrac{x-4}{6}\right) & 4 \le x \le 1 \end{cases}$$

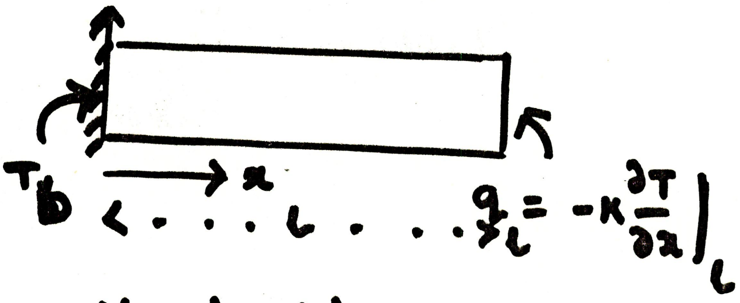
(30

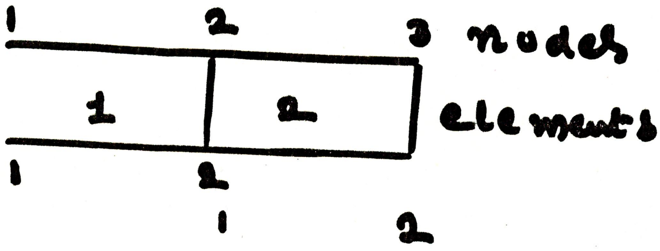
$$q_1^{(1)} = 19\left(\frac{54200}{654} - 100\right) \; W/cm^2. \quad (31$$

Heat Transfer in a Slender Rod

$$K A \frac{d^2 T}{dx^2} - h P (T - T_\infty) = 0 \qquad (1)$$

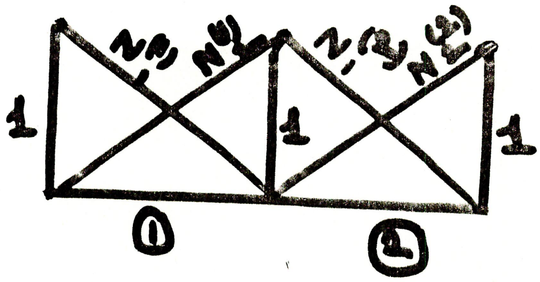
$$x = 0 : T = T_b \; ; \; x = L, \; q = -K \frac{dT}{dx} \qquad (2), (3)$$





Variational formulation

$$I = \int_0^L \frac{1}{2} K A \left(\frac{dT}{dx}\right)^2 dx + \int_0^L \frac{1}{2} h P (T - T_\infty)^2 dx$$
$$+ \int_S q T \, ds \qquad (4)$$



Shape functions
element 1: $N_1^{(1)}$, $N_2^{(1)}$
element 2: $N_1^{(2)}$, $N_2^{(2)}$

Linear shape functions:

$$T = \alpha_1 + \alpha_2 x \qquad (5)$$

$$T_i = \alpha_1 + \alpha_2 x_i \, , \quad T_j = \alpha_1 + \alpha_2 x_j$$

$$\alpha_2 = \frac{T_j - T_i}{x_j - x_i} \, , \quad \alpha_1 = \frac{T_i x_j - T_j x_i}{x_j - x_i} \qquad (6)$$

$$T = \frac{T_i x_j - T_j x_i + (T_j - T_i)x}{x_j - x_i}$$

$$= T_i \frac{x_j - x}{x_j - x_i} + T_j \frac{x - x_i}{x_j - x_i} = T_i N_i + T_j N_j$$

$$= (N_i \quad N_j) \begin{bmatrix} T_i \\ T_j \end{bmatrix}; \quad N_i = \frac{x_i - x}{L} \quad N_j = \frac{x - x_j}{L} \qquad (6)$$

$$\frac{dT}{dx} = \frac{-T_i + T_j}{L} = \frac{1}{L}(-1 \quad 1) \begin{bmatrix} T_i \\ T_j \end{bmatrix} \qquad (7)$$

2 elements (1, 2); each has 2 nodes

Element 1. $I_1 = \int_0^L \left\{ \frac{1}{2} k_1 A_1 \left(\frac{dT}{dx}\right)^2 + \frac{1}{2} h P (T - T_\infty)^2 \right\} dx$

$$= \frac{1}{2} \int_0^{L_1} \left\{ \frac{K_1 A_1}{L_1^2} (-T_i + T_j)^2 + \frac{1}{2} h_1 P_1 (N_i T_i + N_j T_j - T_\infty)^2 \right\} dx \qquad (8)$$

$$I_2 = \int_0^{L_2} \left\{ \frac{1}{2} \frac{K_2 A_2}{L_2^2} (-T_i + T_j)^2 + \frac{1}{2} h_2 P_2 (N_i T_i + N_j T_j - T_\infty)^2 \right\} dx + \int_{S_1} q (N_i T_i + N_j T_j) \, dS \qquad (9)$$

$$I_1 = \frac{1}{2}\frac{K_1 A_1}{L_1}\left(T_1^2 + T_2^2 - 2T_1T_2\right)$$
$$+ \frac{h_1 L_1 P_1}{6}\left(T_1^2 + T_2^2 + T_\infty^2 + T_1T_2 - T_2T_\infty - T_1T_\infty\right) \qquad (10)$$

$$I_2 = \frac{1}{2}\frac{K_2 A_2}{L_2}\left(T_2^2 + T_3^2 - 2T_2T_3\right)$$
$$+ \frac{h_2 L_2 P_2}{6}\left(T_2^2 + T_3^2 + T_\infty^2 + T_2T_3 - T_3T_\infty - T_2T_\infty\right) + q\,T_3\,A \qquad (11)$$

Recall: $T = T_1 N_1 + T_2 N_2$ (eqn. 6)

$$\left.\begin{aligned} I &= I_1 + I_2 \quad \text{and} \\ \frac{\partial I}{\partial T_i} &= 0 \quad (i = 1, 2, 3) \end{aligned}\right\} \qquad (12)$$

$$\frac{\partial I}{\partial T_1} = \frac{K_1 A_1}{L_1}(T_1 - T_2) + \frac{h_1 L_1 P_1}{6}(2T_1 + T_2 - T_\infty) \qquad (13)$$

$$\frac{\partial I}{\partial T_2} = \frac{K_1 A_1}{L_1}(T_2 - T_1) + \frac{h_1 L_1 P_1}{6}(2T_2 + T_1 - T_\infty)$$
$$+ \frac{K_2 A_2}{L_2}(T_2 - T_3) + \frac{h_2 L_2 P_2}{6}(2T_2 + T_3 - T_\infty) \qquad (14)$$

$$\frac{\partial I}{\partial T_3} = \left(\frac{K_2 A_2}{L_2}\right)(T_3 - T_2) + q A \qquad (15)$$

$$\left(\frac{K_1 A_1}{L_1} + \frac{h_1 L_1 P_1}{3}\right) T_1 + \left(\frac{h_1 L_1 P_1}{6} - \frac{K_1 A_1}{L_1}\right) T_2$$

$$= \frac{h_1 L_1 P_1 T_\infty}{6} \qquad (16)$$

$$\left(-\frac{K_1 A_1}{L_1} + \frac{h_1 L_1 P_1}{6}\right) T_1 + \left(\frac{K_1 A_1}{L_1} + \frac{h_1 L_1 P_1}{3}\right.$$

$$\left. + \frac{K_2 A_2}{L_2} + \frac{h_2 L_2 P_2}{3}\right) T_2 + \left(-\frac{K_2 A_2}{L_2} + \frac{h_2 L_2 P_2}{6}\right) T_3$$

$$= \frac{h_1 L_1 P_1}{6} T_\infty + \frac{h_2 L_2 P_2}{6} T_\infty \qquad (17)$$

$$\left(-\frac{K_2 A_2}{L_2} + \frac{h_2 L_2 P_2}{6}\right) T_2 + \left(\frac{K_2 A_2}{L_2} + \frac{h_2 L_2 P_2}{3}\right) T_3$$

$$= -q A + \frac{h_2 L_2 P_2}{6} T_\infty \qquad (18)$$

$$\frac{K_1 A_1}{L_1} = C_1 , \quad \frac{K_2 A_2}{L_2} = C_2 , \quad \frac{h_1 L_1 P_1}{6} = H_1 ,$$

$$\frac{h_2 L_2 P_2}{6} = H_2 \qquad (19)$$

$$[K]\{T\} = \{f\} \qquad (20)$$

$$K = \begin{vmatrix} c_1 + H_1 & -c_1 + H_1 & \\ -c_1 + H_1 & \{(c_1 + 2H_1) + (c_2 + 2H_2)\} & -c_2 + H_2 \\ & -c_2 + H_2 & c_2 + 2H_2 \end{vmatrix} \qquad (21)$$

$$T = (T_1 \quad T_2 \quad T_3)^T \qquad (22)$$

$$f = \{H_1 T_\infty \quad (H_1 + H_2) T_\infty \quad H_2 T_\infty - qA\}^T \qquad (23)$$

Numerical values for the parameters and computer solution.